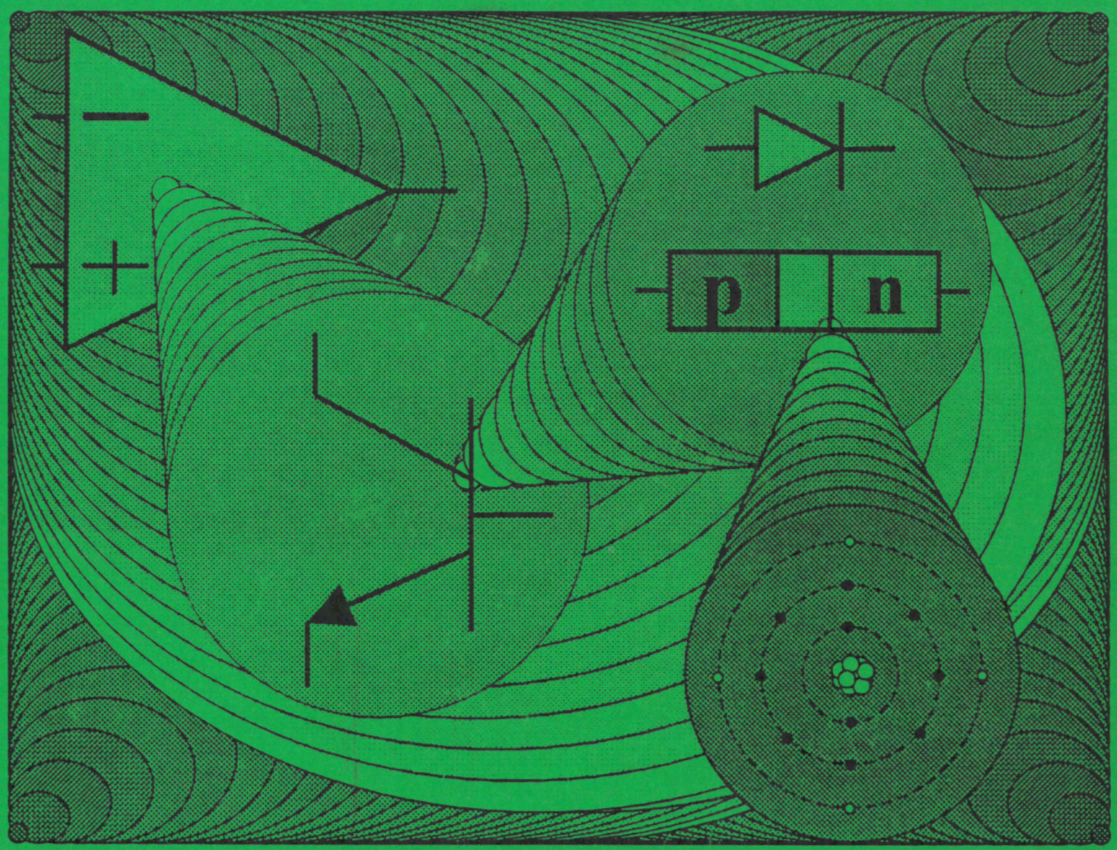
BASIC ELECTRONICS

A Student's Manual

G.M.Tattersfield

(1997 Edition)



ELECTRICAL CIRCUIT ANALYSIS

A Student's Manual

G.M.Tattersfield

(1997 Edition)

INTRODUCTION

This book was written as a manual to accompany a module in Introductory Electronics, given as part of the second-year course in Electrical Engineering at the University of Cape Town. The module is lectured over one semester, and it consists of 21 lectures (plus 2 tests and one closing lecture) each of which is given as a chapter in the following pages. This book is one of a series of student manuals prepared by the same author, and is designed to be read with the similar handbook entitled *Electrical Circuit Analysis.*

It is conventional to teach circuit analysis first, and then to give students a course in basic electronics in the following semester. However, there is increasing pressure to cover the fundamentals of Electrical Engineering rapidly in the early years of a degree course, so as to leave time later for the ever-widening variety of more advanced material. The courses outlined in these two books have been specifically designed at UCT so that the circuits and electronics modules may be given simultaneously. To this end, this *Basic Electronics* handbook has been written with certain particular aims in mind.

These are:

- To provide a course covering fundamental electronics which can be understood readily by a student who is simultaneously studying electrical circuits for the first time.
- To present the material given in each lecture in careful and thorough detail, so that students can read up for lectures in advance, or review a subject after the lecture has been given.
- To introduce the many new ideas and large terminology of electronics without making assumptions of prior knowledge. This is especially important because, in the author's view, a lot of time is often wasted by students seeking answers to simple questions of *fact* before they tackle the salient problems of analysis or design.
- To point out, whenever possible, how an idea fits into the broader field of Electrical Engineering. In particular, the book aims to give the student the feeling of *progressing* towards more advanced topics, and of *using* the mathematics and physics that were studied (often with much effort) in the previous year.

- To encourage students to think for themselves about the concepts encountered in the course and to read further into topics that interest them. However, the book was written in the conviction that very few people will be inspired to read more about something that is confusing from the very start, so the emphasis here is on a very clear statement of the basics.

- Finally, this book aims to communicate the enjoyment that the author has experienced in learning about the subject. Electronics pervades the world about us, and can be appreciated on many planes - practical, recreational, mathematical, etc. - as well as being a fine intellectual discipline in its own right.

To The Student

This book was written to assist you with studying for the Electronics section of Module A of your Second-Year Electrical Engineering course (EEE221W). You will find that each lecture to be given in the course is laid out in detail as a chapter in the book. The best way for you to learn the material in the module is for you to *read each chapter before coming to the corresponding lecture.* This may not always be possible, so you can also use the book to review lectures after they have been given!

It is very important to stress the two things for which this book is *not* a substitute. Firstly, the book will not substitute for lectures! You will always gain more from attending a lecture (like any real-life event) than you will from just reading about it. The course has therefore been carefully designed so that the book will give you the body of the material and the lectures will aim to bring it to life. If you come well-prepared to each lecture, you will enjoy an easy introductory course in an interesting and important field.

Secondly, this is *not* a textbook. There are several dozen excellent and modern texts which cover electronics from every conceivable angle for the student. There is no way for these notes to compare with or to compete with that amount of expertise! You should aim to buy a good electronics textbook, and you can take advice about which would be suitable for you. A good way to make the decision is to browse through some texts in the library, and you ought to make a point of becoming familiar now with the Electrical Engineering section of the UCT library by doing just that.

If you have read the first part of this Introduction, then you will have seen what this book *does* intend to do. It is supposed to be a focus for a 24-lecture course, introducing electronics to you for the first time. You will probably also have a copy of the *Electrical Circuit Analysis* handbook, which accompanies the other 24 lectures of your First Semester module. Together, these 48 lectures should take your understanding of electrical and electronic circuits a long way. You should centre your semester's work around these two books, but you should also be aware that the topics that they cover can be taken much further than you see here. Try to use the books as a launching pad, and not as a pair of crutches!

Please report errors that you may find (or suspect) in the handbook. These can be corrected for the benefit of students taking the course in later years. I hope that you will enjoy reading these notes, and that you will use them to develop a firm foundation for your future Electrical Engineering career.

G.M.T.

Contents

List of Figures

Chapter 1

Theory of Semiconductors

1.1 Electrons and Conduction

When we first consider how electric current can flow through a metallic conductor, we talk in terms of a non-localised "sea" of electrons. These electrons move along the metal if a *voltage* is applied across its ends, and the resulting flow of *charge* is what we call *current*. For historical reasons, *conventional current* is regarded as flowing in the opposite direction to the actual motion of the electrons. The voltage and the conventional current are related in most conductors by *Ohm's Law*, and the linear circuit theory that we study in the other half of this course takes the story from there.

We imagine each metal atom in a conductor to be surrounded by a certain number of electrons: some of these electrons fill inner *shells* around the atomic nucleus, while a few of them occupy the outermost shell, and are called *valence electrons* (see Figure 1.1).

Valence electrons may play one of two possible rôles. Firstly, they may participate in forming bonds between the atom with which they are associated and its neighbours. In this case, they help to determine the exact shape or *lattice structure* of the resulting solid crystal. Secondly, outer electrons may gain enough energy (from light or heat) to break their bonds, and so to become part of the non-localised "sea" of *conduction electrons.*

It is important to realise that both bound electrons and conduction electrons have an associated *energy*, and that there are only certain *discrete* energy levels which either type of electron can possess. Furthermore, each electron type can only have energies within a certain range or *band* of energy values.

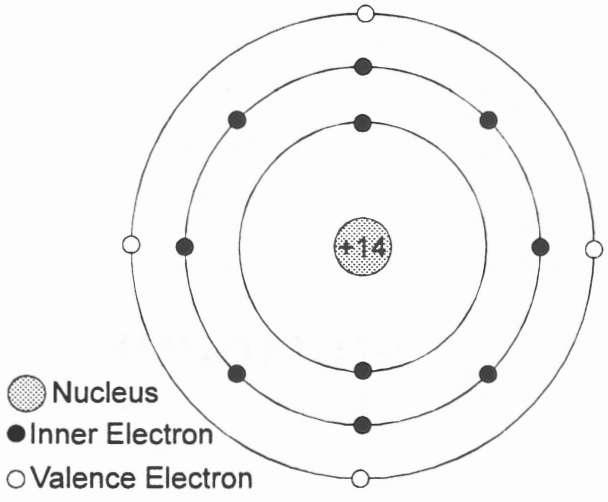


Figure 1.1: The Electrons in the Silicon Atom

Figure 1.2 shows these two bands in the *energy level diagrams* for three different types of material. In Figure 1.2a, the conduction band and the valence band of the material are separated by a wide gap, called the *forbidden band.* This means that it would require a lot of energy to promote an electron that is employed in a bond (i.e. an electron in the valence band) up into the conduction band, where it might become part of a current flow. The energy gap may even be so big that the necessary energy input would destroy the material! Such a material is (at best) a poor conductor of electricity, and we know it as an *insulator.*

By contrast, in Figure 1.2b, the conduction and valence bands overlap, so it is extremely easy to promote a bonding electron to the conduction band. This sort of material is called a *conductor*, and will generally obey Ohm's Law to a first approximation. All metals, and a few non-metals and compounds, fall into this category.

Our main interest in this part of the course lies in materials with energy level diagrams like Figure 1.2c. Here the forbidden band is quite narrow (about 1eV or $1.6 * 10^{-19}$J), and it is possible that an energy input in the form of heat or light could convert the material from an insulator to a conductor. Such materials are known as *semiconductors.*

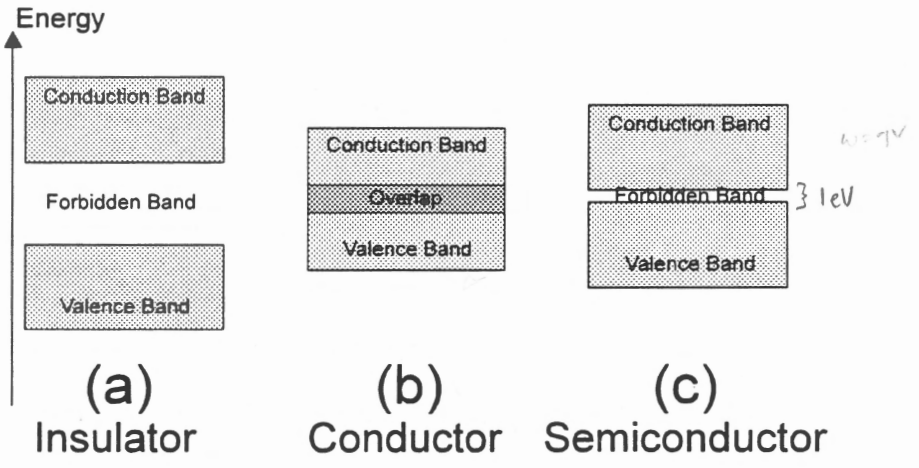


Figure 1.2: Energy Level Diagrams for Different Materials

1.2 Pure Semiconductors

It is found that the forbidden energy gap is wide for materials whose atomic spacing in the crystal is small (e.g. diamond, a good insulator), and conversely that materials with big lattice spacing tend to be conductors (e.g. all metals). It is easy to see that this is because more widely-spaced positive nuclei in the lattice exert (on average) less attractive force on the negatively-charged electrons than closely-packed nuclei would. This means that less energy is required to dislodge electrons from a bonding rôle into a conducting rôle where lattice spacing is higher.

Now, two well-known semiconductors are *silicon* (energy gap 1.1eV) and *germanium* (energy gap 0.67eV). Let us consider them in more detail. Silicon and germanium are in the same group of the periodic table as carbon, and so they tend to form lattices with the same crystal structure as diamond, with each atom bonded to four others by means of *covalent bonds.*

Consider a crystal of pure silicon. At very low temperatures (near absolute zero) the atoms of the crystal hardly vibrate, no covalent bonds are broken, no electrons are promoted to the conduction band, and so the silicon crystal is an insulator. If enough energy (e.g. heat) is applied, then electrons may be promoted from the valence band to the conduction band. Each electron so promoted leaves a *hole* in the valence band, as shown in Figure 1.3.

Evidently, in a pure semiconductor crystal there will be as many holes as

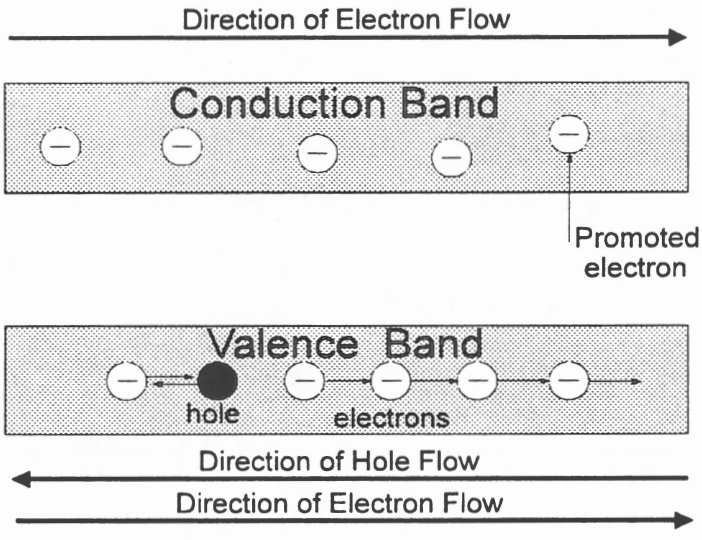


Figure 1.3: Electron and Hole Flow

conduction electrons, and *both* holes and conduction electrons will contribute to the total current flow. Electrons in the valence band may move to fill a nearby hole, thus creating a new, relocated hole. Holes represent the absence of an electron and may be regarded as *positive particles*. Hence, if there is an applied force (i.e. an electric field) acting on all of the electrons in the lattice, then *hole current* in the valence band and the *electron current* in the conduction band are added to find the overall current flow.

Note that in the absence of an externally-applied force, *thermal agitation* of the semiconductor lattice may still cause electron movement (known as *diffusion current*) if, for instance, some mechanism had caused an accumulation of charge at one end of the semiconductor. When a field is applied, however, a different type of electron movement takes place as every electron is accelerated in this field. This is *drift current*, and, by Ohm's Law, drift current can be expected to be proportional to the external field applied.

The *hole concentration* (p in m^{-3}) must equal the conduction electron concentration (n in m^{-3}) in a pure semiconductor. Sometimes a free electron moves close to a hole and is recaptured: the hole disappears and the electron returns to a bonding rôle in the valence band. This is called *recombination*, but since it, too, affects holes and conduction electrons equally, we always

have, for pure (or *intrinsic*) semiconductors:

$$n = p = n_i \tag{1.1}$$

where $n_i = \sqrt{np}$ is called the *intrinsic concentration.* As you might expect, the intrinsic concentration depends on the temperature of the semiconductor crystal, with the actual formula being

$$n_i^2 = AT^3 e^{-E_G/kT} \tag{1.2}$$

where

- A is a constant of proportionality: $5.06 * 10^{43}$ $m^{-6}K^{-3}$ for silicon
- T is the *absolute* temperature in Kelvins
- E_G is the forbidden band energy gap in Joules
- k is Boltzmann's constant: $1.38066 * 10^{-23}$ JK^{-1}

It is interesting to note that at room temperature this formula tells us that there are about $1.45 * 10^{10}$ broken bonds in a cubic centimetre of silicon crystal. Since there are about 10^{23} silicon atoms per cubic centimetre, less than one atom in 10^{12} has a broken bond. This tiny change does not affect the physical, chemical or mechanical properties of the silicon, but it does have an impact on its electrical performance.

1.3 Doped Semiconductors

The electrical properties of silicon are changed dramatically if small amounts of specific impurities are added to the crystal, in a process known as *doping.* The added substance (or *dopant*) often has more than 4 electrons in the outer shell of its atoms, in which case it is called a *donor* and the doped silicon crystal, now having more electrons than holes, is known as an *n-type semiconductor.* In this case, the electrons are called the *majority carriers*, and the holes are the *minority carriers.* Figure 1.4b shows a small section of n-type semiconductor schematically.

If, on the other hand, the dopant has fewer than 4 valence electrons, then it is regarded as donating *holes* to the silicon, and is known as an *acceptor.* The doped crystal is then a *p-type semiconductor*, in which holes are the majority carriers and electrons are the minority carriers. The situation is

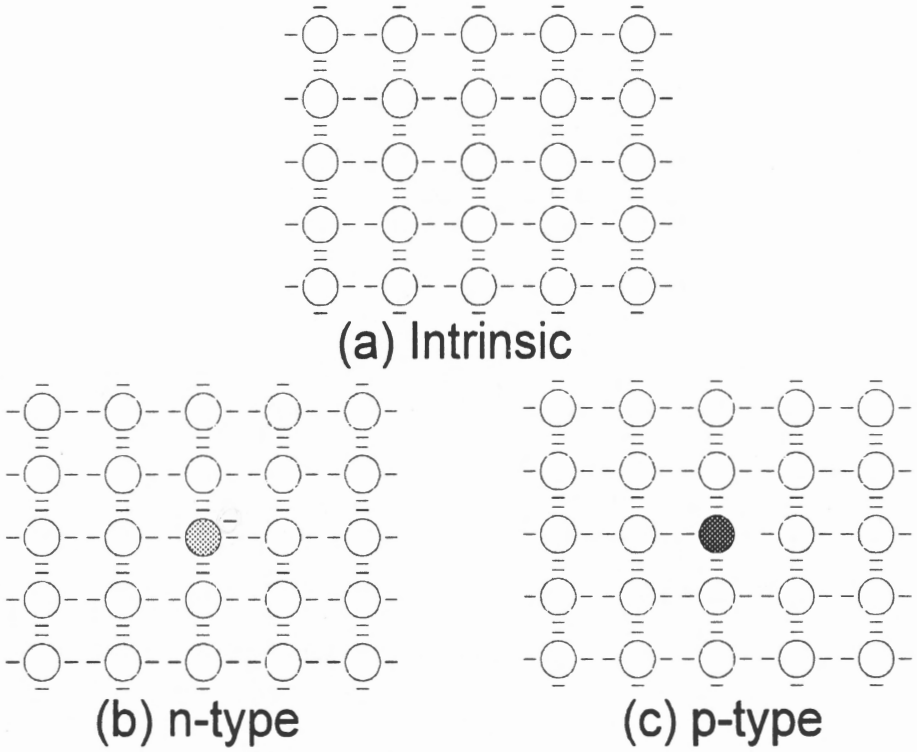


Figure 1.4: Three Different Types of Semiconductor

depicted in Figure 1.4c. Doped crystals are called *extrinsic semiconductors*, while the pure substances are *intrinsic semiconductors* (see Figure 1.4a).

Typical donor impurities for silicon are phosphorus, arsenic and antimony, which all have five valence electrons. The fifth electron does not fit into the lattice, and is very loosely bound, requiring only about 0.05eV to detach it. At room temperature, such an electron would certainly be in the conduction band (although it would *not* give rise to a hole in the valence band, of course). The positively-charged donor atoms are immobile, and their concentration is written N_D.

A common acceptor impurity for silicon is boron, which has three valence electrons. The resulting lattice has a hole for each boron atom in it, which appears to be mobile as electrons take it in turn to fill the vacancy. Again, very little energy is required for this to occur. The acceptor atoms are also immobile, and their concentration is written N_A.

Since holes and immobile donors carry charge $+q$ while electrons and immobile acceptors have charge $-q$, the *local charge density* within part of a silicon crystal can be written

$$\rho = q(p + N_D - n - N_A) \tag{1.3}$$

Ideally, within regions of the semiconductor, there are no pockets of extra charge, so, with $\rho = 0$, we can write

$$n - p = N_D - N_A \tag{1.4}$$

Evidently, N_D and N_A are decided by the manufacturer of the doped semiconductor, but we can see straight away that, for *intrinsic* semiconductors, we have $N_D = N_A = 0$, so $n = p$, as before. For extrinsic (doped) semiconductors

$$n \neq p \neq n_i \tag{1.5}$$

but the *product* of the electron and hole concentrations (i.e. np) is in fact *independent* of the doping, so we continue to write $np = n_i^2$.

Clearly, in n-type materials, the donors far outnumber the acceptors, and the intrinsic carrier concentration will be less than the donor concentration because there are very few holes. Thus

$$N_D >> N_A \qquad N_D >> n_i \qquad p << n \tag{1.6}$$

and so we can get the electron and hole concentrations in terms of the doping concentrations using the expression $n - p = N_D - N_A$:

$$n \approx N_D \qquad p \approx \frac{n_i^2}{N_D} \tag{1.7}$$

In p-type materials, acceptors outnumber donors and there are few free electrons, so

$$N_A >> N_D \qquad N_A >> n_i \qquad n << p \tag{1.8}$$

and so we arrive at

$$p \approx N_A \qquad n \approx \frac{n_i^2}{N_A} \tag{1.9}$$

You should see clearly from these equations that if acceptor atoms are added to silicon crystal at a concentration of only 1 atom to 20 million silicon

atoms (i.e. $5 * 10^{21}$ per m^3), then using the intrinsic concentration of room-temperature silicon which we already know ($1.45 * 10^{16}$ per m^3) we find that in the resulting p-type material the electron concentration is $4.205 * 10^{10}$ per m^3, while the hole concentration is now $5 * 10^{21}$ per m^3 - up from the intrinsic case by a factor of over 300000! In the next section we see that this makes a vast difference to the ability of the silicon material to conduct electricity, even though it is chemically and mechanically hardly changed.

1.4 Effect of Doping on Electric Current

Let us finally look at the effect that electron or hole concentration (*carrier concentration*) has on the movement of charge in a conductor or semiconductor. At any temperature above absolute zero, the lattice will vibrate, and some electrons will move about at random. Although their average *velocity* is zero (there is zero net motion so zero current), their *thermal speed* in ms^{-1} is given by the famous expression

$$v_{th} = \sqrt{\frac{3kT}{m^*}} \tag{1.10}$$

where m^* is the *effective mass* of the carrier, adjusted for the influence of neighbouring atoms (about $1.24 * 10^{-30}$ kg for an electron). You can check that this means, at room temperature, a mean carrier thermal speed of about 10^5 ms^{-1} in silicon.

If a field is now applied, carriers will continue to move with the same thermal speed, but will also *drift* along the field lines, the drift velocity depending on the *field strength* E and on the *carrier mobility* μ. Thus

$$v = \mu E \tag{1.11}$$

Electron mobility for silicon is about 0.135 $m^2V^{-1}s^{-1}$, and hole mobility is approximately 0.045 $m^2V^{-1}s^{-1}$.

Now, to find the resulting electric current, let there be N carriers in a length L of conductor (or semiconductor), as shown in Figure 1.5. If each carrier takes t_o seconds to travel the distance L, then the flow of carriers through the conductor in one second is N/t_o. Hence the current is

$$I = \frac{N}{t_o}q = \frac{Nq(L/t_o)}{L} = \frac{Nqv}{L} \tag{1.12}$$

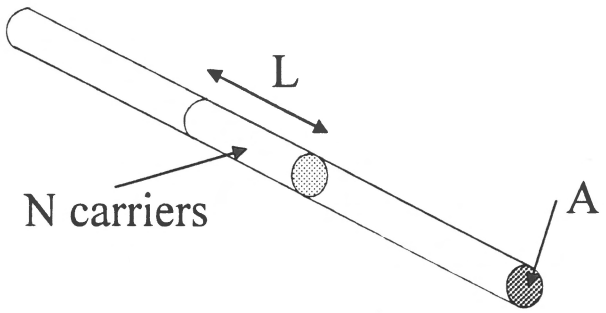


Figure 1.5: A Length of Conducting Material

where q is the charge on an electron and v is the electron velocity. We sometimes talk about *current density* (in Am^{-2}):

$$J = \frac{I}{A} = \frac{Nqv}{LA} \tag{1.13}$$

You should recognise N/LA as just the carrier concentration n, so

$$J = nqv = nq\mu E = \sigma E \tag{1.14}$$

where $\sigma = nq\mu$ is the *conductivity* of the material in $(\Omega m)^{-1}$. Note that we sometimes use *resistivity*, ρ, the reciprocal of conductivity, given in Ωm.

From the above, we now get

$$I = JA = \frac{\sigma EAL}{L} = \frac{\sigma AV}{L} = \frac{V}{R} \tag{1.15}$$

which is, of course, Ohm's Law, since *voltage* $V = EL$ and the *resistance* $R = L/(A\sigma) = \rho/L$.

For holes and electrons in silicon we can now write

$$J_h = qp\mu_h E \qquad\qquad J_e = qn\mu_e E \tag{1.16}$$

so the total drift current density is

$$J = q(p\mu_h + n\mu_e)E \tag{1.17}$$

Do not be outfaced by all these formulæ! Together they represent a body of equations which you will often meet, and with which you should try gradually to become more familiar. The main point at present is that an increase in carrier concentration brought about by doping will lead to a big increase in the drift current in a semiconductor when a field is applied - in other words, conductivity depends very sensitively on doping concentration.

Chapter 2

Semiconductor Diodes

2.1 The pn Junction

A piece of semiconductor that is doped so that one side is p-type and the other side is n-type has some interesting properties. Figure 2.1 shows such a situation, which is called a *pn junction.* Near the junction, free electrons from the n-type material will migrate across to the p-type side and there *recombine* with free holes: likewise, free holes will move from the p-type material to recombine with free electrons in the n-type substance. Both the p-type and n-type materials started with the same number of positive charges in the nuclei of their atoms as negatively-charged electrons (i.e. they were electrically neutral). After the migration of holes and electrons across the junction, however, the p-type material now has extra electrons and therefore net negative charge, while the n-type material has gained some holes and has net positive charge.

You know from basic physics that two nearby charges produce a *field*, and that a *potential difference* exists between any two points in such a field. Because of this field, there comes a time when further electrons on the n-type side are actually repelled from the negatively-charged p-type side, and further holes on the p-type side are repelled by the positively-charged n-type side. The result is a reduction of both types of carrier near the junction, throughout an area known as the *depletion region.* We are left with a field locked inside the semiconductor material, which results in a *potential hill* that any majority carrier must climb if it is to flow across the junction (see Figure 2.1).

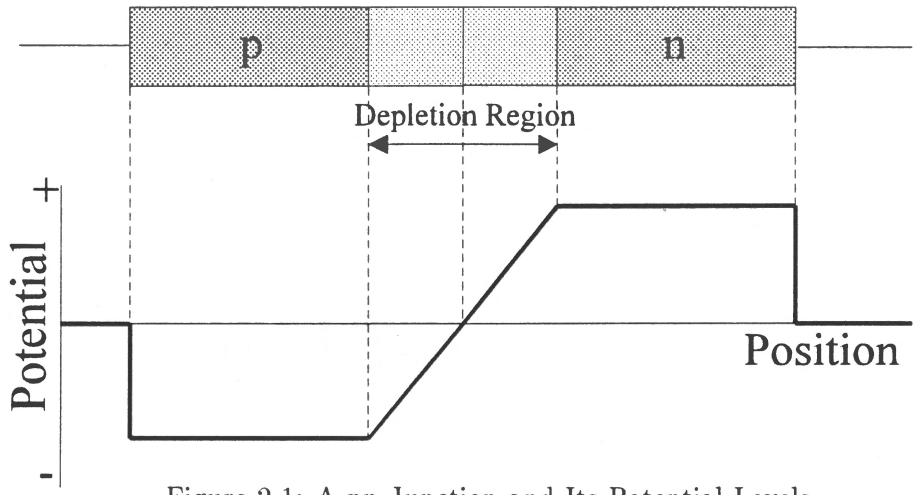


Figure 2.1: A pn Junction and Its Potential Levels

Now, at any moment there are two types of current flow across the junction. The first is the tendency we have already discussed for electrons to move from the n-type side (where they are majority carriers) to the p-type side and for holes to do the opposite. This is known as *diffusion current*, written I_D. The second current arises when thermally-excited minority carriers on both sides arrive at the depletion region, experience the field there, and are swept across the junction by it. Imagine, for example, an electron on the p-side of the depletion layer, with sufficient energy to move to the edge of the layer. There it encounters the field and will accelerate towards the "+" side of the field crossing the junction as it does so. Holes may do the opposite, and together these movements constitute *drift current*, written I_S.

Figure 2.2a shows an open-circuit pn junction, which is in equilibrium with $I_D = I_S$. If we now apply positive potential (a voltage) to the p-side relative to the n-side as shown in Figure 2.2b, the pn junction is said to be *forward-biased.* The newly-imposed field tends to attract majority carriers on each side, so the depletion layer shrinks in size. The potential barrier is thus lowered and so current can flow more readily. In equilibrium, $I_D > I_S$ and a current $I = I_D - I_S$ will flow through the junction as shown. Conversely, if potential of the opposite polarity is applied, as shown in Figure 2.2c, then the junction is *reverse-biased*, the depletion region widens, the potential barrier is raised and current is restricted from crossing the junction.

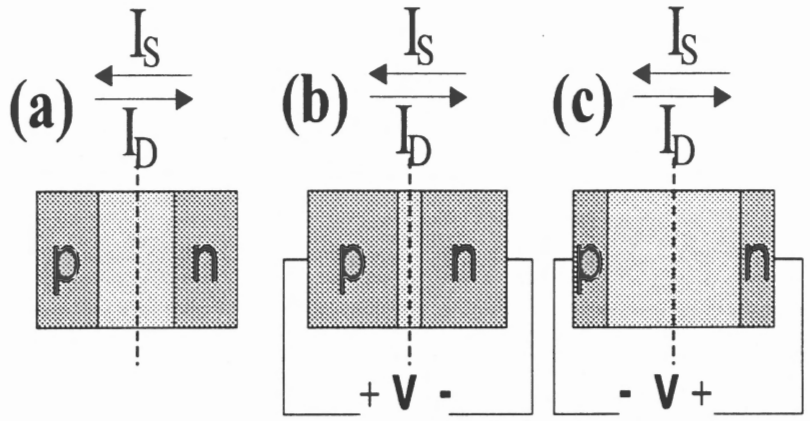


Figure 2.2: Effect of Biasing a pn Junction on the Depletion Region

2.2 Diode i-v Characteristics

We now have the basis for a very useful circuit element. You will know that the simplest circuit element is a *resistor*, and that the current through a resistor varies in response to the voltage across it as shown in Figure 2.3a. The line in the graph is straight, which confirms the resistor as a *linear* element, and the reciprocal of its slope is the *resistance* of the element.

In Figure 2.3b, we see another current-voltage *characteristic curve*, this time describing an element which passes no current (i.e. it acts as an open circuit) if the applied voltage is negative, and which behaves as a short circuit if a positive voltage is applied. You will probably agree that such an element could not exist in practice, but it is a valuable concept in theory, and we call it an *ideal diode*. The symbol we use for the ideal diode in drawing circuits is shown as part of Figure 2.3b. Wherever we see it, we assume that it provides a short circuit to current flowing in the direction of the arrow and an open circuit to current in the opposite direction. The ideal diode i-v characteristic is not a straight line - diodes are *non-linear* elements.

Engineers have been trying to make diodes for nearly 90 years because they have a vast number of applications, but all *practical diodes* give, of course, only an approximation to the ideal performance. Heroic but cumbersome methods ranging from valves to whisker-and-crystals have given way to the pn junction as the most elegant way to realise diode action.

As we have seen, the pn junction allows current flow if forward-biased but

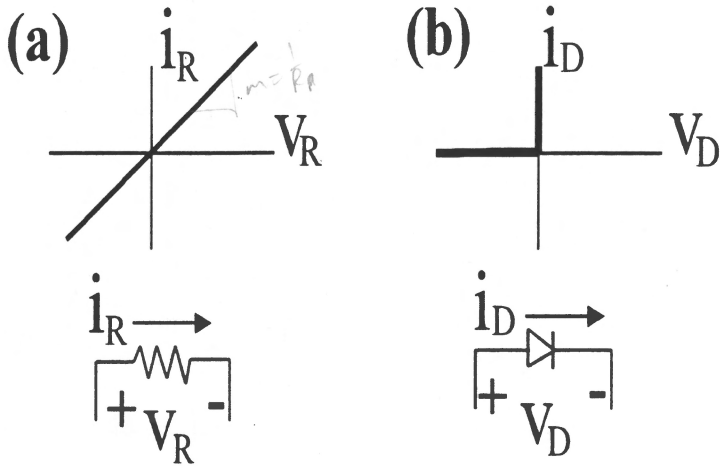


Figure 2.3: Resistor and Diode i-v Characteristics

suppresses it if reverse-biased. The following well-known equation tells how the diode current actually responds to the potential difference applied:

$$i_D = I_o(e^{qv_D/nkT} - 1) \tag{2.1}$$

where

- i_D is the diode current in Amperes
- v_D is the applied potential difference in Volts
- I_o is a constant for a given diode that depends on its doping, geometry and temperature, called the *reverse saturation current*
- n is another empirical constant for a given diode, lying between 1 and 2 and called the *exponential ideality factor*

k, T and q play their familiar rôles as Boltzmann's constant, absolute temperature and electronic charge respectively, and we sometimes see the diode equation simplified by defining $V_T = kT/q$ to give

$$i_D = I_o(e^{v_D/nV_T} - 1) \tag{2.2}$$

At normal operating temperatures (e.g. 300K), and when the diode is forward-biased, the first term in brackets far outweighs the second, so

$$i_D \approx I_o e^{v_D/nV_T} \tag{2.3}$$

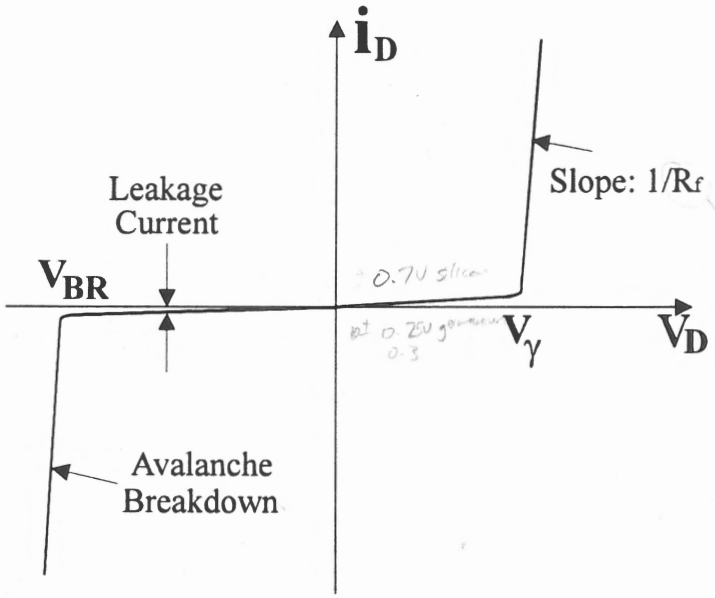


Figure 2.4: Practical Diode Current-Voltage Relationship

2.3 Practical Diodes

The diode i-v curve, derived from the diode equation and closely representing the behaviour of a real diode, is shown in Figure 2.4. You should notice the following features, each of which is a departure from the ideal case:

1. The reciprocal of the slope at any point is the diode resistance, which is a *dynamic variable* that we write as r_d. Under forward or reverse bias, the diode resistance is almost constant, so we denote it as R_f and R_r respectively. We get r_d from the diode equation by observing that

$$\frac{1}{r_d} = \frac{di_D}{dv_D} = \frac{I_o e^{v_D/nV_T}}{nV_T} = \frac{i_D + I_o}{nV_T} \tag{2.4}$$

since $e^{v_D/nV_T} = i_D/I_o + 1$, again by the diode equation. Hence

$$r_d = \frac{nV_T}{i_D + I_o} \approx \frac{nV_T}{i_D} \tag{2.5}$$

because $I_o << i_D$. Note that at room temperature the value of nV_T for silicon is about 26mV when $n = 1$.

2. It requires a minimum voltage V_γ, which is about 0.7V in the case of silicon at room temperature, before appreciable current will flow in the diode. It does *not* flow immediately for any forward voltage as in the ideal diode.

3. A small leakage current flows when the diode is reverse-biased. This is the reverse saturation current, not present in the ideal case.

4. At a large enough reverse voltage, V_{BR}, sometimes known as the *peak inverse voltage* or PIV, the field in the depletion region becomes big enough to produce a dramatic new effect, not predicted by the diode equation. With V_{BR} reverse-biasing the diode, thermally-generated minority carriers at the edge of the depletion region are so strongly accelerated by the field that, upon collision with lattice atoms in the region, they dislodge further electrons (promoting them from a bonding rôle to a conduction rôle and thereby creating a hole as well). These new conduction electrons are also accelerated strongly and can suffer further collisions, soon releasing a cascade of further electrons which move across the pn junction as a large reverse current. The phenomenon is called *avalanche breakdown*, and it is often fatal to the diode because of the excessive heat which it releases.

Finally, the performance of a diode strongly depends on the temperature at which it is operated. The value of the switch-on voltage, V_γ, falls by about 2.0mV per degree Centigrade of temperature rise from its value of about 0.7V at room temperature. This has the effect of shifting the entire characteristic curve to the left in the forward-biased region.

The reverse saturation current, I_o, is also temperature-dependent, and tends to increase by about 7.2% per degree Centigrade of temperature rise. A useful rule of thumb for calculating this effect is

$$I_o(T_2) = I_o(T_1) * 2^{(T_2-T_1)/10} \tag{2.6}$$

Actual values of I_o for a given diode, as well as a lot of other useful diode parameters, are found in the *manufacturer's specifications* for the given device. You would do well to look at a specification for a well-known diode to get the feel of what the manufacturers' sheets look like and to see where to find some of the quantities that we have covered in this chapter.

Chapter 3

Diode Circuit Analysis

3.1 Modelling a Diode

In the other half of this course, we have been calculating currents and voltages in circuits that contain resistors and sources only. How do we analyse circuits that include diodes? There are four approaches to finding the currents and voltages in such a circuit:

- The practical approach - build the circuit and use ammeters and voltmeters to measure the actual results. We will soon be doing this in a laboratory session.
- The computer approach - feed the circuit details into a circuit analysis package such as PSPICE or MICRO-CAP II and call for the required currents or voltages as output. At UCT we use PSPICE in the third year.
- The graphical approach - covered in the next section
- The equations approach - also covered below

In all but the first of these methods we do not actually have the real diode (or other circuit components) in front of us, and instead we make do with certain theoretical assumptions about how diodes perform, which we call our *model.* A diode model can be quite simple, if we are only interested in mimicking the broader features of diode action, or it may be extremely complex if we want a highly accurate reflection of the component's behaviour. It takes many lines of code and numerous parameters to model a diode in PSPICE.

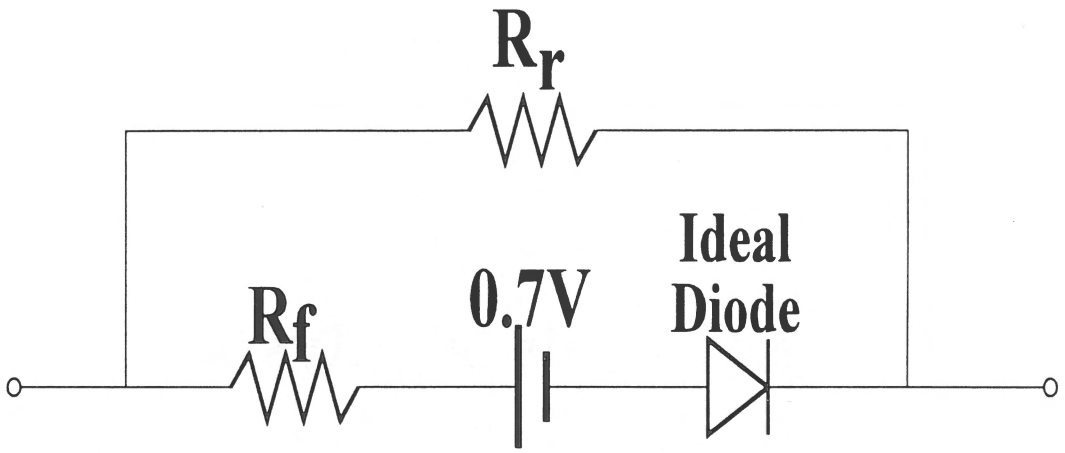


Figure 3.1: The dc Diode Model

Let us look at a straightforward but very useful practical diode model, shown in Figure 3.1. The practical diode is somewhat like an ideal diode, but it needs 0.7V of forward bias before it "switches on". If you could connect a 0.7V voltage source in series with an ideal diode, a little thought will confirm that this combination will perform in the same way. Under reverse-bias, a realistic diode exhibits a resistance R_r (usually several MΩ), and even under forward bias there is some resistance, R_f, due to the diode contacts and the actual silicon material (usually less than 50 Ω). We include these in our model in Figure 3.1, where you should be able to verify that, under reverse bias, the resistance is indeed R_r, while under forward bias our model has resistance

$$R_f // R_r \approx R_f \tag{3.1}$$

Now that we understand (i.e. believe in) the model, we can put it to work for us. Suppose that we want to find the voltage v_o in Figure 3.2a. Firstly we must decide if the circuit diode is forward- or reverse-biased. Initially, this might have to be a guess, but if it turns out from the final full analysis of the circuit that the diode current or voltage is in the opposite direction to our guess, then we shall know that we guessed wrong! Looking at the 10V source, it seems reasonable here that the diode is *forward-biased.*

Next we must replace the diode in the circuit with our diode model. This is shown in Figure 3.2b. Notice that R_r is not shown because (as we have seen) R_f dominates under forward-bias conditions. The forward bias also

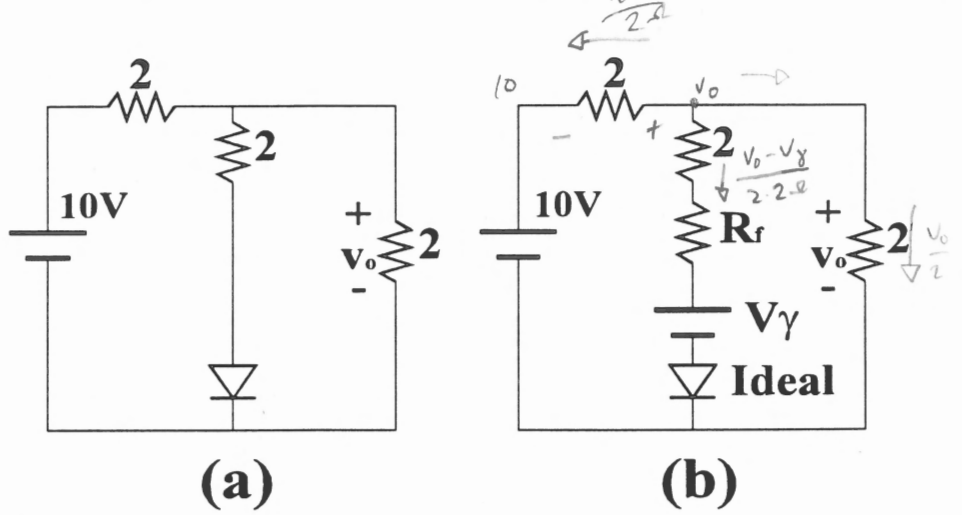


Figure 3.2: Applying the dc Diode Model to Circuit Analysis

means that the ideal diode is short-circuited, so we are now left with a simple circuit analysis problem in which v_o can be found by applying Kirchhoff's Current Law at a single node:

$$\frac{v_o - 10}{2} + \frac{v_o - V_\gamma}{2 + R_f} + \frac{v_o}{2} = 0 \tag{3.2}$$

You can verify that, with the typical values of $V_\gamma = 0.7$V and $R_f = 0.2\Omega$, we will get a final value of $v_o = 3.66$V. Can you see that, if the diode had been connected the other way around, this method would rapidly give us $v_o = 5.0$V?

3.2 Diode Load Lines

There is an instructive and elegant graphical way to analyse circuits, which is often useful when a diode is present. Consider the circuit of Figure 3.3a, and suppose that we want to find v_o and i_o. You can do it easily, of course, using your circuit analysis techniques, but let us use the graphical approach instead. The resistor with voltage v_o across it and current i_o through it is called the *load resistor*, and Ohm's Law tells us that, although we don't yet know v_o or i_o, they are related by

$$v_o = (1)i_o \qquad or \qquad i_o = v_o \tag{3.3}$$

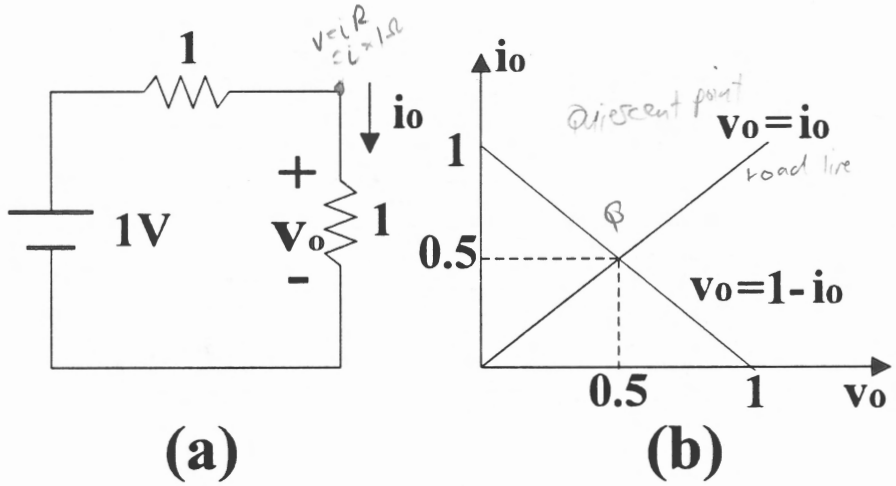


Figure 3.3: Load Line Technique in a Resistive Circuit

We can plot this as a line (called a *load line*) on a graph of i against v. Now, from the voltage source and the other resistor it is easy to see that

$$v_o = 1 - (1)i_o \qquad so \qquad i_o = 1 - v_o \tag{3.4}$$

and, again, we can plot this on the graph (see Figure 3.3b). Solving the two equations above simultaneously amounts to finding the point of intersection of the two lines on the graph - this is known as the *operating point.* You can read off the result ($V_o = 0.5$V and $I_o = 0.5$A) that we knew anyway.

The graphical technique is of more value in situations like the one in Figure 3.4a, where we are interested in finding v_D and i_D. We could, of course, use a diode model approach as in the previous section, but, in this example, *the source may vary over time* and our diode model is not sophisticated enough to handle that very well. Assuming that the source voltage is always positive, the load line for the diode is simply the forward-biased section of its i-v characteristic curve, while the other line is given by the KVL equation

$$v_D = V_S - Ri_D \qquad so \qquad i_D = V_S/R - v_D/R \tag{3.5}$$

This line meets the i_D axis at V_S/R and meets the v_D axis at V_S. It is plotted, together with the load line, in Figure 3.4b, and the operating point could be used to find the diode voltage and current when the source voltage is a constant V_S. One would have to get the actual diode curve from the manufacturers' notes on the device to get numerical answers.

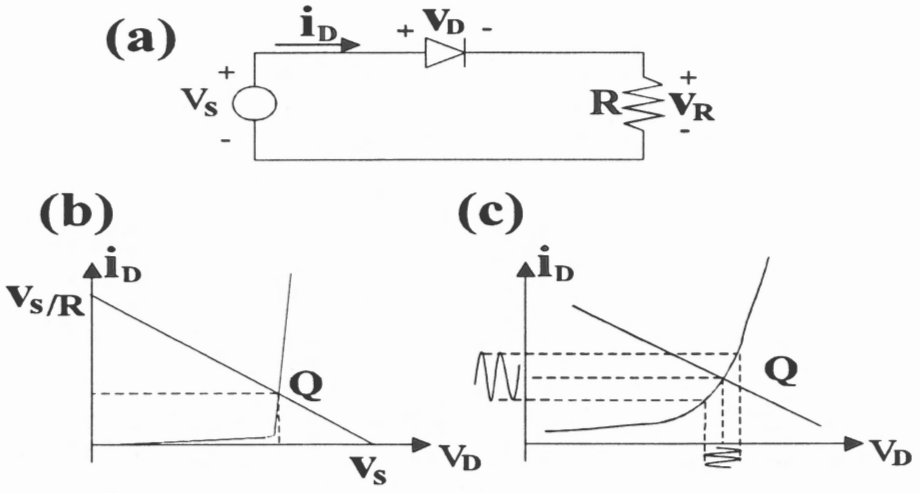


Figure 3.4: Graphical Load Line Technique in a Diode Circuit

Now, suppose that the source voltage actually varies around V_S as an average value, but has a sinusoidal variation v_s added, as shown in Figure 3.4c. So long as $v_s = 0$, the diode is at its *quiescent operating point* (or *Q-point*) on the graph; but any variation in the source voltage will cause the operating point to shift along the diode curve, and the diode voltage and current to vary over time in a similar way. Try to imagine how the diode voltage and current would respond to a much bigger swing in the source voltage. Can you see that their shapes will be *distorted* because the diode line is not straight? You will meet this phenomenon often and become very aware of its significance as your studies in electronics progress.

3.3 Using Diode Equations

It is a good idea to be able to use a few simple relationships to get quick answers to questions about diode operation, rather than having to draw a graph or program a computer or build a circuit every time! As an example, suppose that we want to know the diode current i_D in Figure 3.4a and we are aware that the source voltage is $1.1 + 0.1\sin(1000t)$V - made up of a "dc" component, V_S, and a varying part, v_s. We are also told some diode parameters: $nV_T = 40$mV and $V_\gamma = 0.7$V. Finally, $R = 100\Omega$.

KVL quickly gives us $V_S = V_\gamma + I_D R$, so

$$I_D = \frac{V_S - V_\gamma}{R} = \frac{1.1 - 0.7}{100} = 4mA \tag{3.6}$$

So we already have the quiescent (dc) value for the diode current! For *varying* input voltages, the diode exhibits a *dynamic resistance*, r_d, which we saw in the previous chapter is given by

$$r_d \approx \frac{nV_T}{i_D} \tag{3.7}$$

Hence, at diode currents around 4mA, the diode resistance will be about (40mV)/(4mA)= 10Ω. Then, by KVL again, we can write the time-varying voltage as

$$v_s = r_d i_d + R i_d \tag{3.8}$$

so

$$i_d = \frac{v_s}{r_d + R} = \frac{0.1\sin(1000t)}{110} = 0.91\sin(1000t)mA \tag{3.9}$$

Thus, the diode current (both its dc and time-varying parts) is

$$i_d = 4 + 0.91\sin(1000t)mA \tag{3.10}$$

3.4 Power Handling by Diodes

There are many types of diode available for use in a wide variety of situations, and it is important to use an appropriate device in any application. One of the ways of rating diodes is according to their power-handling capability, which varies over several orders of magnitude for different devices. At any instant, the power produced by a diode is given by

$$p_D = v_D i_D \tag{3.11}$$

This power is expressed as heat, which must be dissipated away, and this is often done, if the heating is severe, by mounting the diode on a metal *heatsink*. If the diode is unable to release the heat it produces, it could fail to function, but in diodes carrying light currents this is rarely a problem.

Chapter 4

Zener Diodes and Regulators

4.1 Zener Diodes

Let us now move our attention from the forward-biased region to the reverse-biased region of the diode characteristic curve. A certain special class of diodes, called *zener diodes*, are designed to have an extremely steep breakdown characteristic, and to be able to survive the avalanche that occurs at breakdown, up to some maximum current called I_{Zmax}.

Over and above the avalanche breakdown that occurs in normal diodes, in zener diodes there is a second mechanism that promotes avalanche. A sufficiently strong field can break the bonds holding an electron in place, thereby promoting it to the conduction band (and leaving a hole), without any collision taking place. This second effect, known as *zener breakdown*, greatly increases the number of electrons in the avalanche, and so steepens the breakdown section of the diode curve. Furthermore, by carefully controlling the doping in the zener diode, an exact reverse voltage (the *zener voltage*) at which zener breakdown occurs can be ensured.

We have, then, a device which will break down at a specified point and which can survive a reverse current of about ten times the leakage current that it passed just before breakdown. In other words, it will not be destroyed by a power dissipation up to

$$P_{Zmax} = V_Z I_{Zmax} \approx 10 V_Z I_{Zmin} \tag{4.1}$$

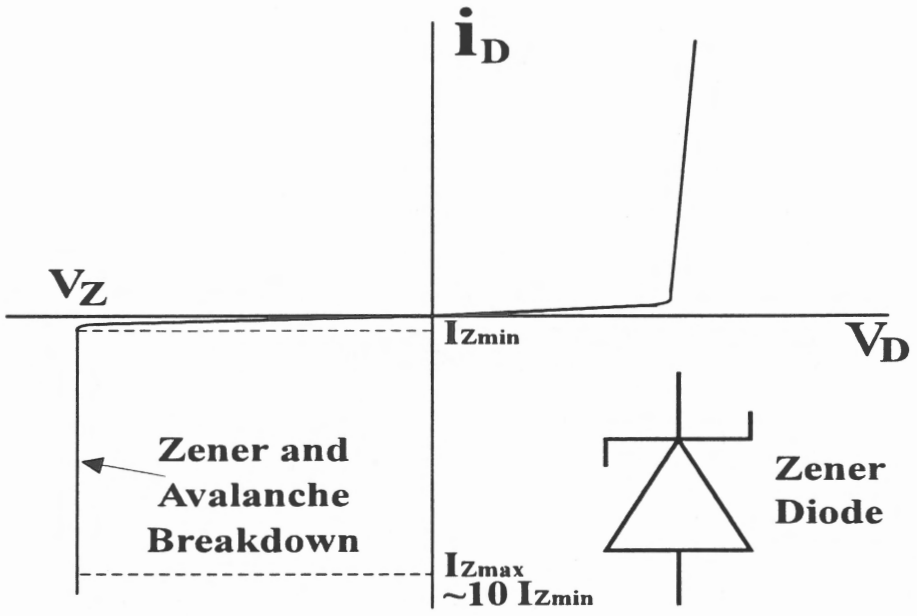


Figure 4.1: The Zener Diode Characteristic Curve and Symbol

Figure 4.1 shows the characteristic curve for an ideal zener diode, and also the circuit symbol used to denote that a given diode has got these special zener characteristics.

4.2 The Zener Regulator

We often use constant voltage sources in circuit theory, without stopping to think how such a device might be realised in practice. The zener diode offers a way to build a simple circuit that provides a reliable and steady voltage at its output (within certain limits). The big advantage of this is that you can supply a load whose resistance *changes*, perhaps as it heats up or as parts of the load switch on and off, without the voltage across the load fluctuating, even though the current drawn by the load alters from moment to moment. This would be so much better than using an unstable voltage supply, which might ultimately damage the load or cause it to perform poorly.

Have a look at Figure 4.2, in which the load is presented as a variable resistor, R_L. The (unstable) voltage source v_s is being *regulated* by the *zener regulator* comprising the resistor R_i and the zener diode. Notice, first, that

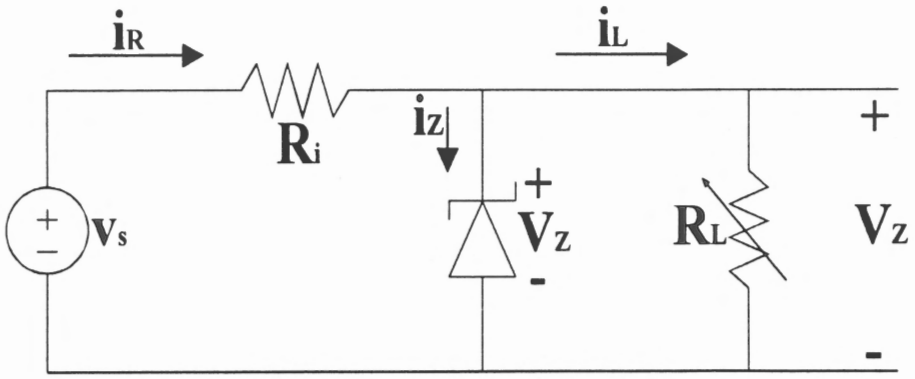


Figure 4.2: A Zener Regulator Supplies a Load

the zener diode is reverse-biased and is, in fact, operating in its breakdown region. This means, in terms of Figure 4.1, that the voltage across the diode is constant at V_Z, while the current through the diode is above I_{Zmin} and (hopefully) below I_{Zmax}.

Now, if R_i were a small resistance, then it seems reasonable that the current through it, i_R, would be large; and, since V_Z is constant and therefore i_L is constant (at any instant), this means that the zener current i_Z would be large as well. Perhaps it might exceed I_{Zmax} and thus be *too* large. Conversely, if R_i were a big resistance, then i_Z might be so small that the diode would no longer be in the breakdown region of the zener curve, and the regulator would fail to supply voltage V_Z to the load. Clearly, our choice of R_i is critical to whether the zener regulator works: indeed, when we speak of *designing* a zener regulator, all we mean is choosing the right resistor for R_i to do the job. We shall now see how to do this.

In Figure 4.2, KVL gives us

$$v_s - i_R R_i - V_Z = 0 \tag{4.2}$$

and since KCL gives us $i_R = i_Z + i_L$, we have

$$v_s - (i_Z + i_L)R_i - V_Z = 0 \qquad so \qquad R_i = \frac{v_s - V_Z}{i_Z + i_L} \tag{4.3}$$

This can easily be re-written to make i_Z the subject of the formula:

$$i_Z = \frac{v_s - V_Z}{R_i} - i_L \tag{4.4}$$

Now, v_s varies (because the source is a bit unreliable) and so does i_L (because the load resistance fluctuates). Obviously the diode current, i_Z, would be at its minimum (I_{Zmin}) if i_L were at its highest value (I_{Lmax}) and v_s were at its lowest (V_{Smin}). Likewise, the diode current $i_Z = I_{Zmax}$ if $i_L = I_{Lmin}$ and $v_s = V_{Smax}$. Our equations above demand that we always have

$$R_i = \frac{v_s - V_Z}{i_Z + i_L} \tag{4.5}$$

so we can write

$$R_i = \frac{V_{Smin} - V_Z}{I_{Zmin} + I_{Lmax}} = \frac{V_{Smax} - V_Z}{I_{Zmax} + I_{Lmin}} \tag{4.6}$$

and therefore

$$(V_{Smin} - V_Z)(I_{Zmax} + I_{Lmin}) = (V_{Smax} - V_Z)(I_{Zmin} + I_{Lmax}) \tag{4.7}$$

We know that we want the zener diode's current to operate between I_{Zmin} and I_{Zmax}, where

$$I_{Zmin} = 0.1 I_{Zmax} \tag{4.8}$$

so we can re-write the above equation for I_{Zmax} as follows:

$$I_{Zmax} = \frac{I_{Lmin}(V_Z - V_{Smin}) + I_{Lmax}(V_{Smax} - V_Z)}{V_{Smin} - 0.9V_Z - 0.1V_{Smax}} \tag{4.9}$$

Since, in a practical situation, we would know the range of fluctuation of the source voltage (V_{Smin} to V_{Smax}) and load current (I_{Lmin} to I_{Lmax}), we have all that we need to find I_{Zmax} and I_{Zmin}, from which we then easily get R_i.

The process appears rather complex at first, but after you have done a couple of examples you will see that it is really very easy, and in the laboratory you will be able to check how effective the simple zener regulator is. As a first example, let us design a 10V zener regulator for an electronic digital clock, given that the current drawn by the clock varies from 100mA to 200mA as its output display changes, and that the source voltage may range from 14V to 20V. From our equations we get

$$I_{Zmax} = \frac{0.1(10 - 14) + 0.2(20 - 10)}{14 - 0.9(10) - 0.1(20)} = 0.533A \tag{4.10}$$

(N.B. if I_{Zmax} comes out negative here, it means that the required zener regulator is *not possible*, given the specified voltage and current ranges.)

From the above,

$$R_i = \frac{V_{Smax} - V_Z}{I_{Zmax} + I_{Lmin}} = \frac{20 - 10}{0.533 + 0.1} = 15.8\Omega \tag{4.11}$$

We must also specify the *power rating* of the resistor, or else we might choose a 15.8Ω resistor that simply cannot fulfil the required task. The greatest current that might ever flow in R_i is

$$I_{Rmax} = I_{Zmax} + I_{Lmin} = 0.633A \tag{4.12}$$

Similarly, the greatest voltage that might ever exist across R_i is

$$V_{Rmax} = V_{Smax} - V_Z = 20 - 10 = 10V \tag{4.13}$$

So, the greatest power that might be dissipated in the resistor is

$$P_{Rmax} = V_{Rmax} I_{Rmax} = 6.33W \tag{4.14}$$

To be quite safe, we would probably choose a 10W, 15.8Ω resistor for R_i. Note that we must also specify the zener's power rating. This is simply

$$P_Z = V_Z I_{Zmax} = 10 * 0.533 = 5.33W \tag{4.15}$$

Any 10V zener with a power rating above 5.33W would be satisfactory.

4.3 Percent Regulation

A real zener diode does not have the absolutely vertical breakdown characteristic that we have assumed so far, but in fact has some tendency for the reverse current to increase in response to an increase in reverse voltage. Although this tendency is slight, it means that we should *model* a real zener diode as having a small resistance (typically a few Ohms) in series with an ideal zener diode, to account for this behaviour. Figure 4.3 shows this simple practical zener diode model, and a realistic characteristic curve.

The slope in the breakdown section means that, as the zener current varies from I_{Zmin} to I_{Zmax}, we actually get some variation in the zener voltage. Hence the regulator is not perfect. In the example above, let us suppose that the zener diode resistance is $R_Z = 2\Omega$. We know that the zener current varies from 0.0533A to 0.533A, and the zener voltage is ideally 10V. You should see from Figure 4.3 that the output voltage will vary between

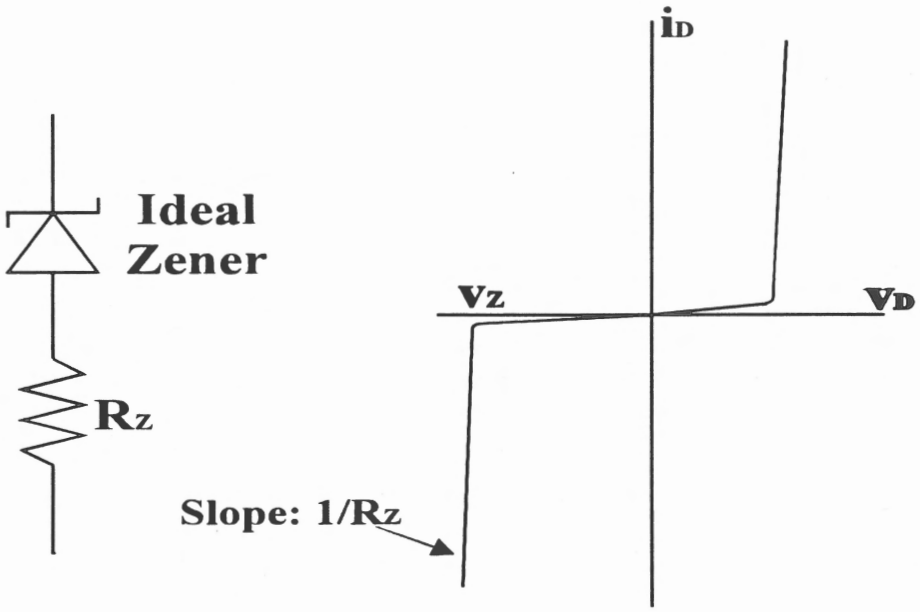


Figure 4.3: A Practical Zener Diode and Its Characteristic Curve

$$V_{omin} = 10 + (0.0533 * 2) = 10.11\text{V}$$

and

$$V_{omax} = 10 + (0.533 * 2) = 11.07\text{V}$$

We express this variability in the voltage output of the regulator as the *percent regulation*, defined as

$$\%Reg = \frac{V_{omax} - V_{omin}}{V_Z} = \frac{11.07 - 10.11}{10} = 9.6\% \qquad (4.16)$$

This is not a very good value for a regulator, but it could be reduced by changing R_i so that I_{Zmax} was not so big, and there would then be less current variation in the zener diode.

Chapter 5

Other Types of Diode

5.1 Light Emitting Diodes

When a diode is forward-biased, we have seen how the width of the depletion region is reduced, leaving only a small potential gradient for carriers to climb in order to cross the pn junction. When current flows across the junction, electrons move to the hole-rich p-type side, where some of them *recombine* with holes. As we know, recombination involves the electron falling from the high-energy conduction band back down into the valence band. To conserve energy, this must involve some sort of energy release; and, in the case of silicon or germanium diodes, this release is in the form of heat (a conducting diode can be expected to warm up). In other words, energy is released at *infra-red* frequencies.

Diodes can be fabricated from *gallium arsenide*, rather than from silicon. One of the many remarkable properties of this material is that recombination of electrons and holes in it gives rise to photon emission centered in the near infra-red. If some *gallium phosphide* is added to the gallium arsenide, then the diode actually emits visible light; and further fine-tuning of the doping can determine the precise colour of the peak radiation emitted. Thus, it is possible to obtain diodes which, when forward-biased, actually glow red, orange, yellow, green or even blue as they conduct current.

The *light-emitting diode* (*LED*) has become a commonplace of daily life over the last 20 years, and its applications, principally in instrumental displays but also in telecommunications and optical electronics, are numerous.

Figure 5.1 shows an LED connected to a simple power supply and in series with a current-limiting resistor. Note that the LED is always connected so

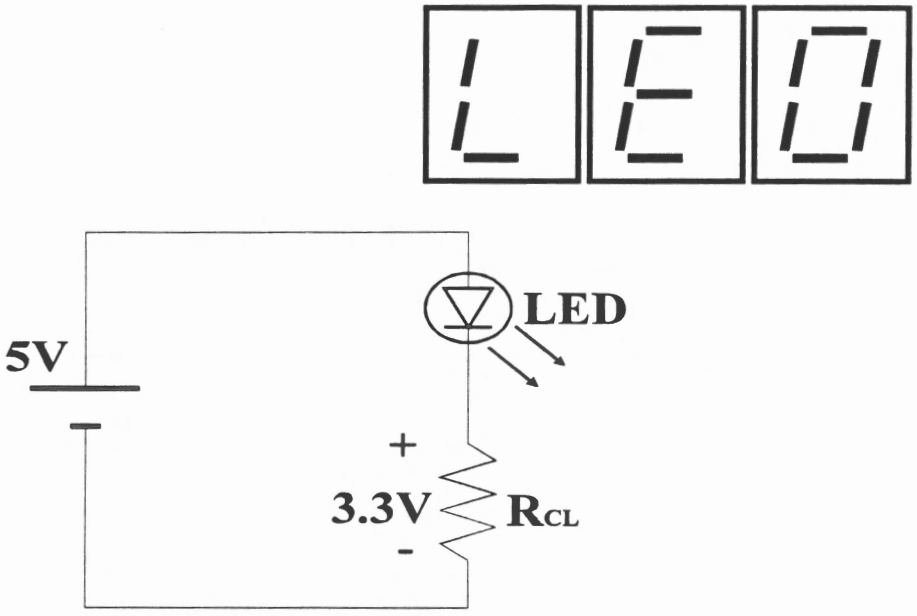


Figure 5.1: A Circuit Containing a Light Emitting Diode

as to be forward-biased. The resistor is necessary because, as we know, the forward resistance of diodes is small. Without the current-limiting resistor, the current that would flow through the LED would probably destroy it. Let us see how to decide the size of this resistor.

When forward-biased, an LED typically has a forward voltage drop of about 1.7V, and can safely pass a current of around 10mA. Hence, in Figure 5.1, the voltage to be dropped across the resistor R_{CL} is $5 - 1.7 = 3.3$V. To set the current at 10mA, we therefore require $R_{CL} = 330\Omega$, and this is in fact a commonly-used resistor value in 5V LED displays.

5.2 Photodiodes

A *photodiode* achieves the opposite of an LED, in converting light energy into electrical current. Figure 5.2a shows a photodiode connected in a very simple circuit, and you should notice immediately that it is *reverse-biased.*

You may recall from earlier lectures that conduction occurs in diodes when energy (heat or light) is added, promoting electrons from the valence band to the conduction band. Under reverse-bias, minority carriers that diffuse

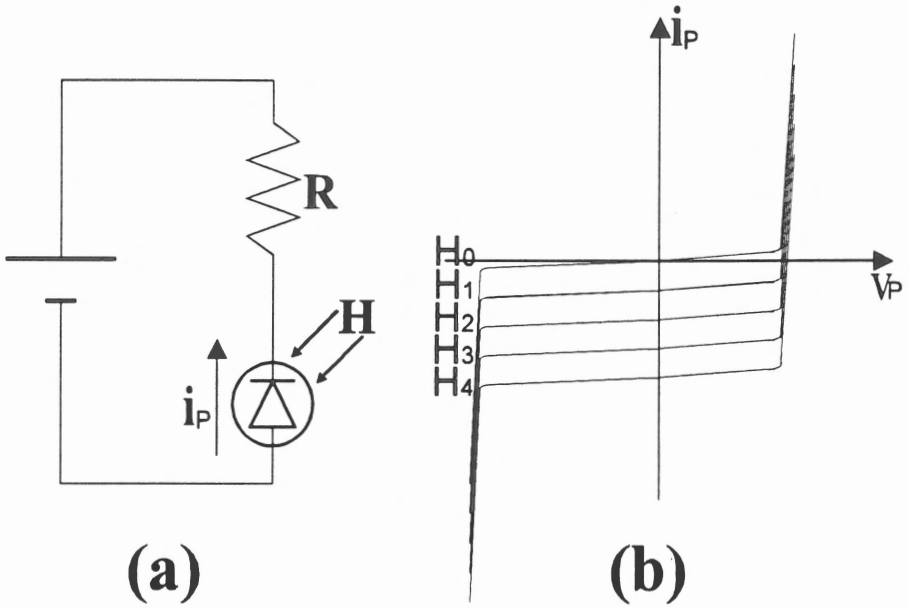


Figure 5.2: A Photodiode and Its Characteristic Curves

to the edge of the depletion region may be swept across the junction by the field. This establishes a small leakage current, known as the reverse saturation current, whose size is determined by several factors, including the amount of energy being injected. Figure 5.2b shows the effect upon the leakage current of increasing the incident energy in a carefully-doped silicon photodiode.

At low light intensity, the photodiode has the characteristic curve H_o, but if the light falling on the photodiode increases, more conduction electrons are then produced and the reverse leakage current rises, making the characteristic curve migrate towards H_4. So long as the reverse voltage does not exceed the breakdown threshold, the photodiode therefore behaves as a constant current generator, with the value of the current depending on the incident light intensity. You could attach an ammeter in series with the photodiode and get a simple (uncalibrated) light meter.

There are many applications of photodiodes ranging from daylight detectors and interrupted-beam-type burglar alarms to uses in telecommunications and electronics. Have a look inside your computer's "mouse" and see if you can find both LEDs and photodiodes at work in it!

learn one (npn) and then apply reverse to (pnp)

Chapter 6

Modelling the BJT

6.1 The Bipolar Junction Transistor

Hopefully you have already been impressed by some of the things that can be achieved by bringing p-type and n-type material into contact. So far, we have considered only the effects of careful doping around a *single* junction. Another wide range of possibilities is opened up if *two* pn junctions are positioned close to one another. This might take the form of two pieces of n-type material separated by a thin p-type slice, or alternatively there could be two p-type pieces with a little n-type sliver between them. In either case, when there are two pn junctions placed back to back, the overall device is termed a *bipolar junction transistor* (or *BJT*).

Figure 6.1 illustrates both types of BJT and gives the symbol for each. For ease of reference, we call them the *npn transistor* and the *pnp transistor*, and you will observe that they are both three-terminal devices. In both cases, the three layers of material are named the *emitter*, the *base* and the *collector*, with the base being the thin layer in the middle.

Let us start by explaining how the npn transistor works (the pnp device works on similar principles, but things get confusing if we study more than one at a time!). In the npn BJT, the emitter is a heavily-doped, medium-sized layer that is therefore rich in electrons, which it is designed to emit or *inject* into nearby material. Right next to the emitter, the base is a medium-doped, very thin p-type layer, ideal for passing electrons through. On the other side of the base (at the receiving end) is the collector, which is a lightly-doped, rather large layer whose rôle is to collect the electrons as they come through from the base.

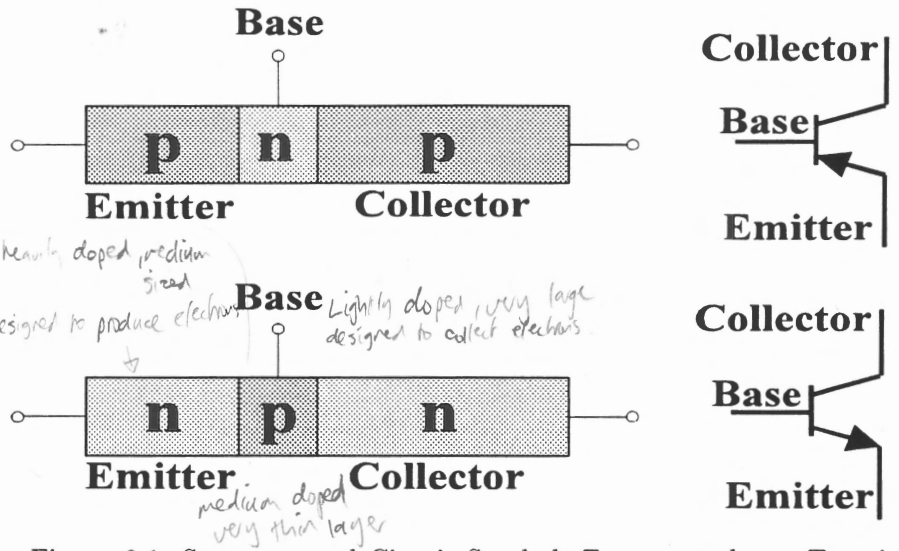


Figure 6.1: Structure and Circuit Symbols For pnp and npn Transistors

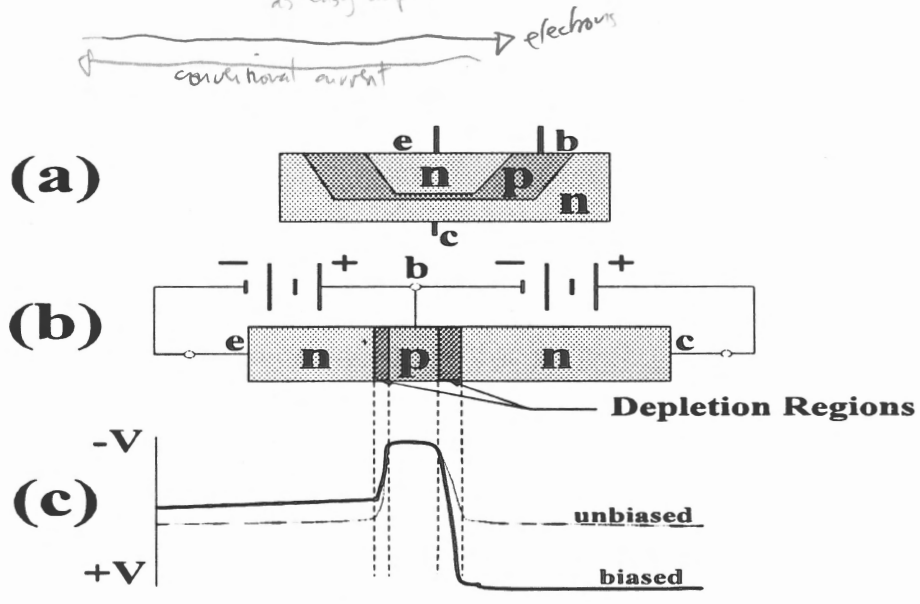


Figure 6.2: The Potential Hill in a Biased npn Transistor

Figure 6.2(a) shows how common "planar type" npn transistors are actually constructed, using n-type and p-type material. Note how the collector-base junction is made to be physically bigger than the emitter-base junction to ensure that the electrons which the emitter produces will be captured by the collector, even if they spread out considerably as they cross the base.

Examining Figure 6.2(b), you see a *biased* transistor. The emitter-base junction is forward-biased and the base-collector junction is reverse-biased, and in Figure 6.2(b) this is achieved by using two separate voltage sources. Can you see how it might be done with a single voltage source and two resistors, or with a voltage source and a single *variable* resistor? Try to sketch these methods for yourself. Figure 6.2(c) shows the potential graph for the BJT. As you might expect, the graph goes slightly uphill at the forward-biased junction and considerably downhill at the reverse-biased pn interface.

In an *unbiased* npn transistor, the size of the potential step at the emitter-base junction is approximately equal (but opposite) to the size of the potential step at the base-collector junction. This means that if we graph the potential as a function of position across the device, we obtain something like the dotted potential line shown in Figure 6.2(c). If this were the case, you can confirm that an emitter electron (which has negative charge) would encounter a considerable potential barrier to its movement across the emitter-base junction, but that an electron in the base would "see" a "downhill" potential slope in the direction of the collector, due to the field caused by the charges piled up on either side of the base-collector depletion region. It helps to remember that the electron has negative charge, and so its potential energy is *raised* if it moves to a more negatively-charged region, and *lowered* if it moves to a region of more positive charge.

Now consider the *biased* npn transistor of Figure 6.2(b). In this case, the forward biasing on the emitter-base junction reduces the potential step from emitter to base, while the base-collector junction is reverse-biased and so the potential step across that junction increases in height because of the biasing. This is shown in the solid graph of the potential in Figure 6.2(b). The sizes of the biasing voltages can be chosen so that any electron leaving the emitter "sees" only a small potential hill at the forward-biased emitter-base junction. Some electrons in the emitter may well possess the thermal energy required to climb such a small hill, and can therefore move up into the base region. After this, they diffuse through the p-type base region (which is very thin) as far as the base-collector junction, where they are rapidly swept down the potential slope to the collector.

If the biasing were altered to give less forward bias on the emitter-base junction, then the potential hill at that junction would grow higher. This would mean that fewer electrons would be injected from the emitter to the base, with the result that the overall transistor current (of electrons from emitter to collector, or of conventional current from collector to emitter) would be reduced, and perhaps even cut off entirely. Conversely, if the base-biasing voltage were increased, then the uphill potential slope would be flattened and many more electrons could climb it and proceed through the transistor. We therefore have a kind of electronic tap or switch - simply adjust the bias voltage on the base, and you control the flow of electrons from the emitter to the collector. Thus, we say that the base-emitter voltage controls the conventional current flowing from collector to emitter in a transistor.

You may be wondering why electrons injected into the base from the emitter don't simply leave the base and return to the voltage source. A few electrons (about 1 or 2 percent of those leaving the emitter) do diffuse to the base connector. However, the main bulk of the injected electrons cross the base rapidly (because it is so thin) and then come under the influence of the field at the second junction, and are pulled to the collector before they have a chance either to recombine with a hole or to reach the base connector.

Figure 6.3a shows in detail the currents that flow in an npn transistor. Notice the small current i_B that enters the BJT at the base. This corresponds to the electrons which escape at the base connection, and is expressed in the diagram as a flow of *holes* from the p-type base back to the emitter. Notice also that, as the base-collector junction is reverse-biased, it exhibits a small leakage current generally written I_{CBO}. The voltages on the base and collector (with respect to earth) are V_{BB} and V_{CC} respectively, and the transistor is protected from receiving dangerous current levels by the resistors R_B and R_C.

You will sometimes see a BJT described as being controlled by the small current i_B that flows into its base. If base current i_B did not flow, you can see in terms of Figure 6.3(a) that there would be a gradual build-up of electrons in the base (due to recombination there with holes). This would give the base a greater negative charge, and would therefore raise the potential hill at the emitter-base junction, cutting off electron flow through to the collector. Hence, the main electron flow from emitter to collector (and so also the conventional current flow in the opposite direction) *requires* that there be a small current i_B into the base. In this sense, the transistor current can be controlled by controlling i_B.

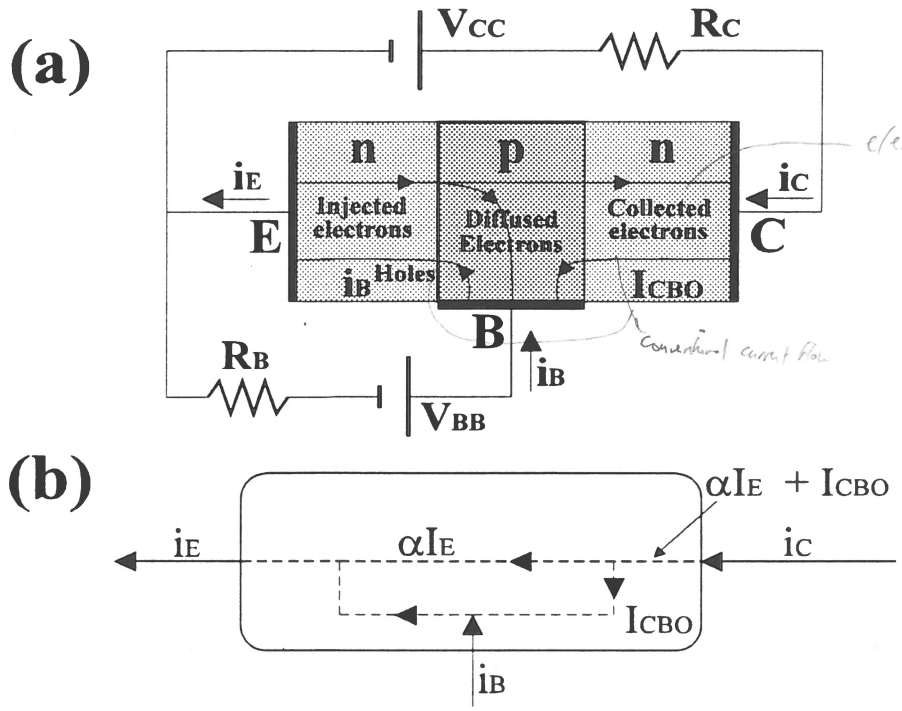


Figure 6.3: The Currents in an npn Transistor

6.2 Transistor Current Relationships

We are now in a position to write some useful equations for the npn transistor, based upon our visual summary of the current flows in Figure 6.3b. Evidently, KCL for the BJT as a whole gives

$$i_E = i_C + i_B \tag{6.1}$$

(Notice how i_B and i_C are defined positively as flowing *into* the transistor, but i_E is regarded as a current which flows *out* of it.)

Within the transistor, evidently part of the emitter current comes directly from the collector (we write this fraction as αi_E) while part comes in the form of holes from the base. Thus

$$i_C = \alpha i_E + I_{CBO} \qquad so \qquad i_E = \frac{i_C - I_{CBO}}{\alpha} \tag{6.2}$$

We may note in passing that α is the rate at which collector current changes in response to emitter current variation (under given biasing) and is known

as the *common-base current gain.* "Gain" seems a slightly inappropriate term, since α is less than 1 - usually between 0.9 and 0.999. The two equations above now give

$$i_E = \alpha i_E + I_{CBO} + i_B \qquad so \qquad i_B = i_E(1-\alpha) - I_{CBO} \tag{6.3}$$

We can substitute for i_E in this last equation to get

$$i_B = \frac{(i_C - I_{CBO})(1-\alpha)}{\alpha} - I_{CBO} = \frac{(1-\alpha)i_C}{\alpha} - \frac{I_{CBO}}{\alpha} \tag{6.4}$$

Since $1/\alpha \approx 1$, we write this as

$$i_B \approx \frac{1-\alpha}{\alpha} i_C - I_{CBO} \tag{6.5}$$

We now *define* another gain factor, called β, to be the rate of change of collector current as base current varies. So

$$\beta = \frac{\alpha}{1-\alpha} \qquad and \qquad i_B \approx \frac{i_C}{\beta} - I_{CBO} \tag{6.6}$$

You can measure β directly for a given transistor using some kinds of multimeter, and we will do this in the laboratory. It lies typically between 10 and 1000, and its full name is the *large-signal amplification factor.* Sometimes you see it written h_{fe}. It is important to realise that in practice no two transistors are identical, and values of β will vary even among two transistors of the same batch.

The equations above form the basis of a good understanding of currents in a transistor, and you should try to be familiar with them. Notice, as well, that if we choose to ignore the small leakage current I_{CBO}, then we get

$$i_C \approx \beta i_B \tag{6.7}$$

from our last equation. Finally, ignoring I_{CBO} *and* assuming $\alpha = 1$ in Equation 6.2, we arrive at the ultimate simplification:

$$i_C \approx i_E \tag{6.8}$$

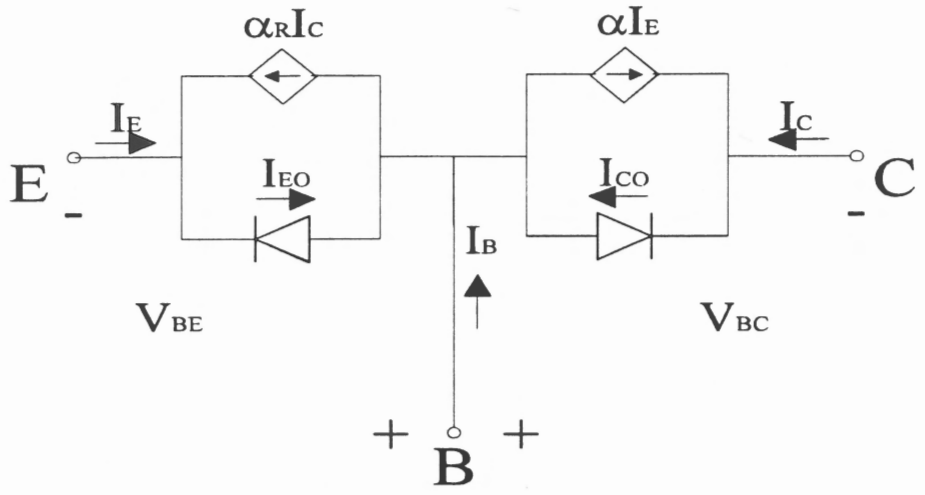


Figure 6.4: The Ebers-Moll Model of an npn Transistor

6.3 A Transistor Model

We used models of both the standard diode and the zener diode to fill out our understanding of how they perform. Likewise, several transistor models exist, which are used in electronic circuit analysis, and which are of varying degrees of complexity. At this stage we will look at just one very well-known BJT model, called the *Ebers-Moll model.*

Models such as the Ebers-Moll model are of value to engineers who produce computer simulations for large circuits using programs such as PSPICE. Our aim here is simply to get a feeling for how the action of the BJT can usefully be simplified. One confusing aspect is that it is accepted practice when discussing the Ebers-Moll model to show all of the terminal current flowing *into* the device. This makes it very hard to visualise the true directions of currents with reference to the diagrams in this section. Nevertheless, we shall see how the model confirms some important basic transistor equations. At this stage, study the explanation of the model and try to convince yourself that the model is *appropriate* (i.e. that it works in the same way as we understand transistors to work, and might therefore be useful to us). As before, we consider only the npn transistor, because the procedure for pnp transistors is somewhat similar.

The Ebers-Moll model is really based on two observations derived from Figure 6.3 and from the discussion of the last section:

- Current from the base to the collector is made up from some fraction of the emitter current (αi_E) *less* whatever current may leak backwards through the base-collector junction. Our best guess at what this reverse saturation current might be is to equate it to the reverse collector current that would flow if $i_E = 0$, which we write I_{CO} Thus

$$i_{BtoC} = \alpha i_E - I_{CO} \tag{6.9}$$

- Current from the base to the emitter is made up from some (other) fraction of the collector current ($\alpha_R i_C$) *less* any current leading backwards to the base through the base-emitter junction. Again, we write this reverse saturation current as though $i_C = 0$, and give it as the constant I_{EO}. Thus

$$i_{BtoE} = \alpha_R i_C - I_{EO} \tag{6.10}$$

  The *reverse gain*, α_R, is usually below 0.5 (lower than α because of the steeper hill that collector electrons need to climb to reach the base).

These equations are summed up in the circuit of Figure 6.4, which uses dependent current sources to model the behaviour of the npn transistor. You will probably need to study the model rather closely before you agree fully with (i.e. understand) it, but it will repay your time! Note that, since Figure 6.4 depicts a *particular* state of bias, we write all of the currents as *constants*, using capital letters.

If the BJT model of Figure 6.4 is to be of any use at all, it must tally with the BJT equations of the previous section. Let us see if it does. Firstly, we remember that the equations referred to a BJT whose base-emitter junction was forward-biased and whose base-collector junction was reverse-biased. This means that between base and collector we can be sure of a current of $\alpha I_E - I_{CO}$ and, since I_{CO} is relatively small, we can just write this as a single current source αI_E (see Figures 6.5a and 6.5b). Meanwhile, between the base and the emitter, the forward-biased diode short-circuits the current source delivering $\alpha_R I_C$ and we can ignore I_{EO} which is also very small. The full simplification for the biased BJT is shown in Figure 6.5b.

Applying KCL to Figure 6.5b now gives us

$$I_B + I_E = \alpha I_E = -I_C \tag{6.11}$$

and so

$$I_E = \frac{I_B}{\alpha - 1} \tag{6.12}$$

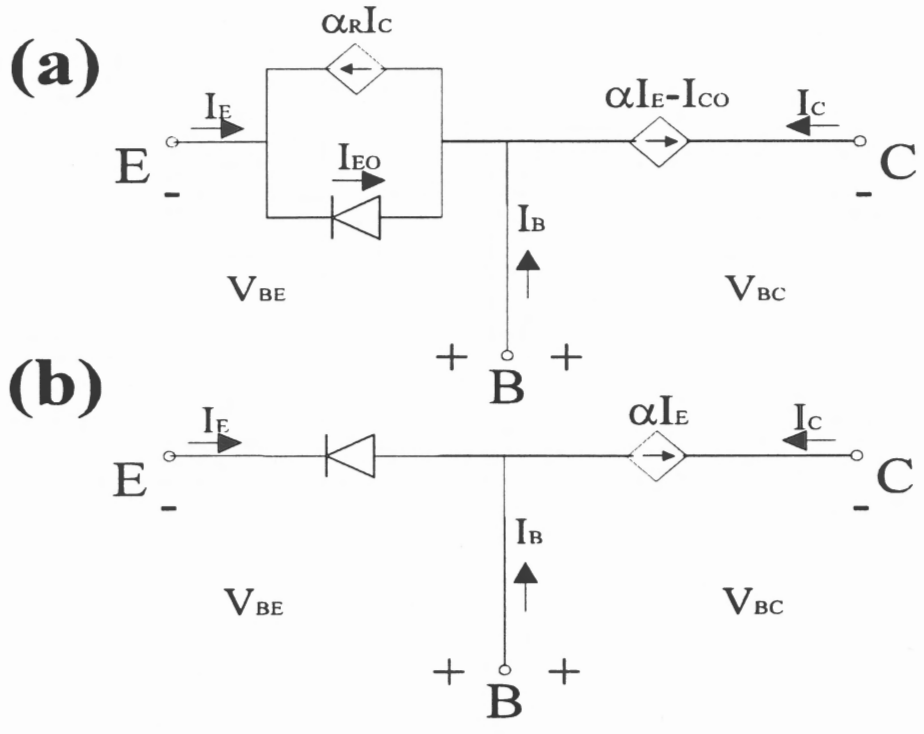


Figure 6.5: Simplifications of the Ebers-Moll Model

giving

$$I_C = \frac{\alpha}{1-\alpha} I_B = \beta I_B \tag{6.13}$$

just as we had before. This is also an approximation, of course, because of the simplifications that we made to the model. So, we can proceed with some confidence in using the Ebers-Moll model and its reduced versions. You will frequently find yourself mentally replacing a transistor with a diode and current source, to see some aspect of a circuit's behaviour clearly. You can now be sure that this is a justified simplification. Keep an eye open for other transistor models (e.g. the *hybrid-π* or the *h-parameter* models). Every time that you have a new way to investigate a different facet of BJT behaviour, it will deepen your appreciation of these extremely versatile devices.

Chapter 7

BJT Characteristic Curves

7.1 BJT Circuit Configurations

Like diodes, transistors are *non-linear* devices. By this we mean that currents or voltages which we regard as outputs from a transistor in a circuit are not always proportional to other currents or voltages, which we regard as inputs. There are several ways that transistors are to be found connected in circuits, and these are shown in Figure 7.1.

The most frequently-used configuration, given in Figure 7.1a, is the *common-emitter* (*CE*) connection. Here we regard the input as a voltage on the base with respect to the ground (or sometimes as a current to the base), and the output is taken as the voltage of the collector, again with respect to ground. Notice the source voltages V_{BB} and V_{CC} which are chosen to ensure that the BJT is *biased* as we discussed in the last chapter, with the base-emitter junction forward-biased and the base-collector junction reverse-biased. We will be returning many times to the common-emitter circuit, and this and other courses will devote considerable attention to its many properties.

Figure 7.1b shows a BJT connected in the *common-collector* configuration, which you will also see described as an *emitter-follower* circuit. As with the CE circuit, the input is a voltage on the base, but the output of the emitter-follower is the voltage at the emitter. Both voltages are measured with respect to ground. Try to memorise both this and the CE circuit, and take good note of their differences and similarities.

Figure 7.1c shows the less usual *common-base* configuration, where the input is the emitter voltage and the output is the voltage at the collector. This is, in fact, the circuit that we used to derive the Ebers-Moll model of the BJT.

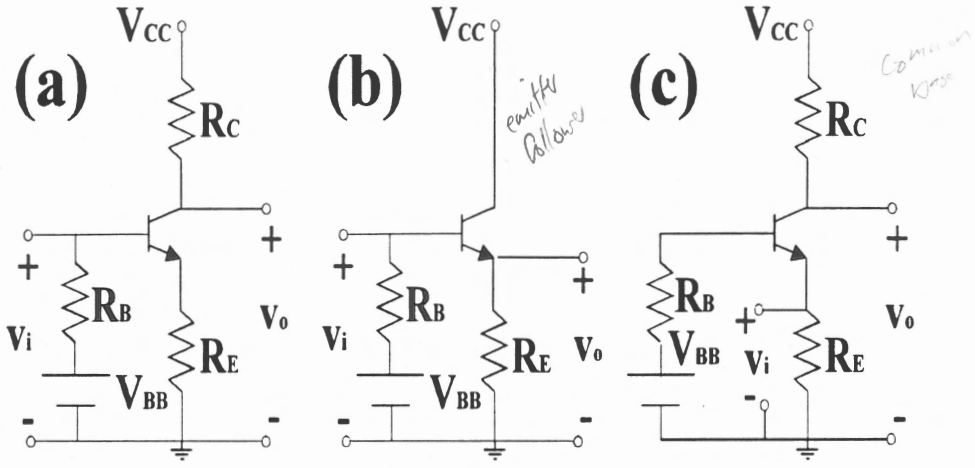


Figure 7.1: The Main BJT Circuit Configurations

7.2 Emitter-Base Characteristics

There is another consequence of the fact that transistors are non-linear, and that is that the best way to define how they will operate is to do so graphically, using *characteristic curves*, as we did with diodes. We should emphasise straight away that there are *many* ways of graphing the performance of a BJT, and if you look up the manufacturer's notes on a given device, you may be rather bewildered by all the graphs that they give. To keep it simple, as we aim here to get a better understanding of how transistors operate, we are only going to look at *two* very important graphs. As ever, we will deal only with the npn transistor, because once you have got the main ideas, understanding the pnp transistor follows fairly easily.

Suppose you connected up an npn BJT as shown in Figure 7.2a, so that the voltage between the collector and the emitter, v_{CE}, is held constant but the base-emitter input voltage, v_{BE}, is allowed to vary. If we regarded the output as being the current from the emitter, i_E, then we would find that it responded to varying input something like what is shown in the graph of Figure 7.2b. This result should not surprise us, because all we are doing is looking at the current passing through the base-emitter junction as we vary the base-emitter voltage, so naturally we obtain a simple forward-biased diode characteristic.

It is, however, a very important result because it means that, so long as we hold the collector voltage reasonably stable with respect to the emitter

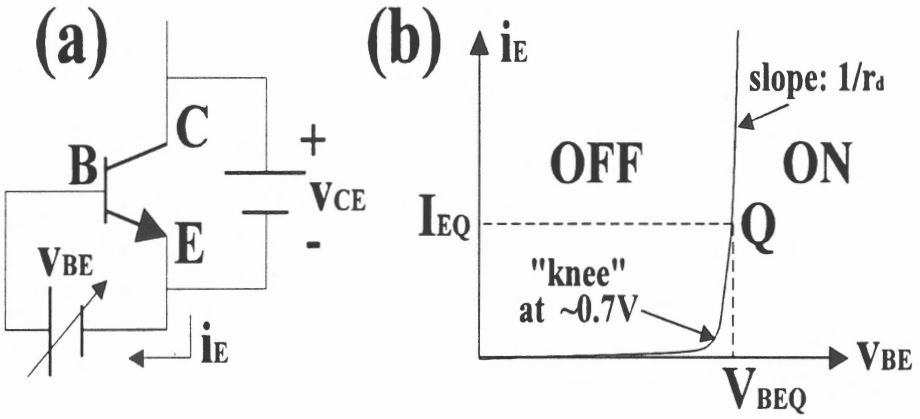


Figure 7.2: The Emitter-Base Characteristic of an npn BJT

voltage, we can use the BJT as a *switch*. For low values of v_{BE}, almost no emitter current flows, but the emitter current rises sharply after a certain base-emitter voltage is reached. In silicon transistors, this occurs around $v_{BE} = 0.7\text{V}$, and is known as the *cut-off point* or the "knee voltage". In germanium transistors, the knee voltage is about 0.2V.

As with diodes, we often talk about the average or stable point on its characteristic curve at which a transistor is operating as its *Q-point*, which has coordinates (V_{BEQ}, I_{EQ}) on the graph. Now suppose that a BJT is connected as in Figure 7.2a (notice we have left out all the resistors for simplicity) and is operating with $v_{BE} = V_{BEQ}$ and $i_E = I_{EQ}$ as shown in Figure 7.2b. The transistor will exhibit *dynamic resistance* between its collector and its emitter, which we can easily see (from an equation that we derived for *any* diode junction) to be

$$r_d \approx \frac{nV_T}{i_E} \tag{7.1}$$

Assuming a silicon junction with $n \approx 1$, this gives us

$$r_d \approx \frac{0.026}{i_{EQ}} \approx \frac{0.026}{I_{CQ}} \tag{7.2}$$

since $I_{CQ} \approx I_{EQ}$ for any transistor. So, not only do we have a switch, but we can also quote its resistance when switched on, the reciprocal of which is, of course, the slope of the "on" section of the BJT emitter-base characteristic.

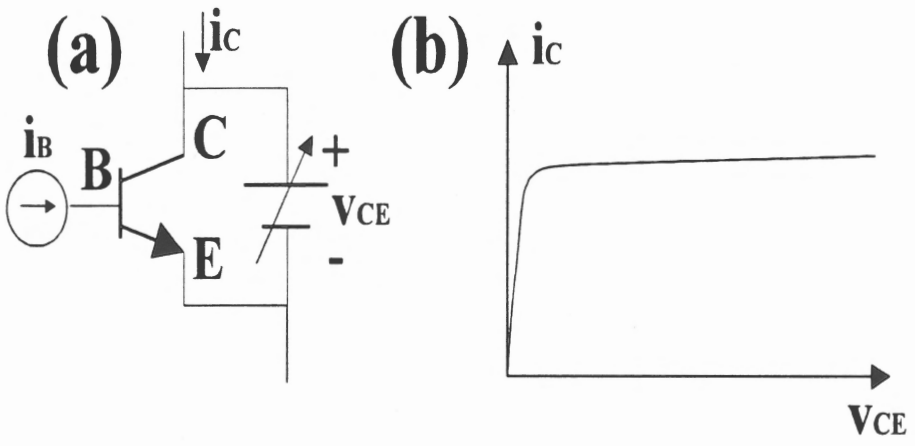


Figure 7.3: The Collector-Emitter Characteristic of an npn BJT

7.3 Collector-Emitter Characteristics

For our second transistor curve, we consider the response of the collector current to changes in the voltage between the collector and the emitter, at a given base current. We imagine connecting an npn BJT as shown in Figure 7.3a and slowly increasing the input voltage, v_{CE}, from zero. We know already that collector current, like emitter current, depends primarily upon the BJT being properly biased. It is not a function of the overall voltage, v_{CE}, across the BJT; as long as there is a reasonable total voltage drop, the transistor will be able to pass current. So we find that the graph of the collector current response in Figure 7.3b is almost flat over most of its length (i.e. varying v_{CE} has virtually no effect), except when v_{CE} is very small, whereupon i_C falls towards zero.

Figure 7.3b is the response at a single value of base current, but we get a much better picture of how the BJT performs if we draw a set of such curves for a range of base current values, as in Figure 7.4. Now, if we know the base current into a transistor, i_B, and we know *one* of the collector current or the collector-emitter voltage, we can use this graph to find the other. In fact, knowing any two of the parameters i_c, v_{CE} and i_B, you can now always find the third.

You will notice that, over much of the graph, the curves are very straight and evenly-separated, so the three parameters are in good, proportional relationships. This area is known as the *linear range of operation* of the

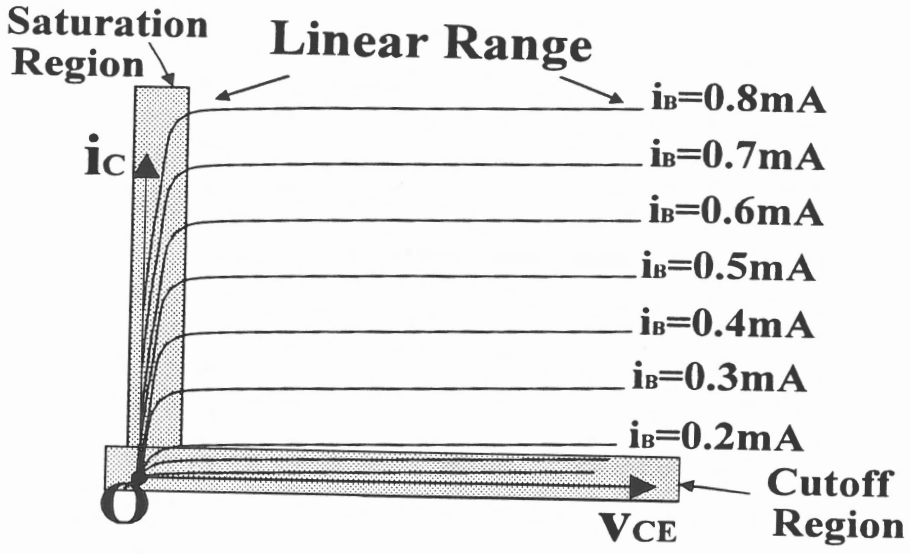


Figure 7.4: A Set of CE Transistor Characteristic Curves

BJT. In many applications, a designer will wish a transistor to be working in this area. When v_{CE} is small, on the other hand, all of the lines are curved and the proportionality breaks down. This is called the *saturation region.* Likewise, at small values of i_B, i_C approaches zero in a non-linear manner, so designers will often avoid using a BJT in this part of the graph, known as the *cut-off region.*

We can now use the collector-emitter curves to solve a simple transistor circuit (i.e. find all of the voltages and currents in it). Consider the circuit of Figure 7.5a, in which a transistor drives a current through resistors R_C and R_E. We need to know where the BJT's operating point lies, and we are given only that the base current $i_B = 0.5\text{mA}$. As with diode circuits, this is going to need a *load line*, obtained by looking at the resistors through which current will flow (R_C and R_E). In the main current loop. KVL gives us

$$V_{CC} - i_C R_C - v_{CE} - i_E R_E = 0 \tag{7.3}$$

and, since $i_E \approx i_C$, this becomes

$$V_{CC} = i_C(R_C + R_E) + v_{CE} \tag{7.4}$$

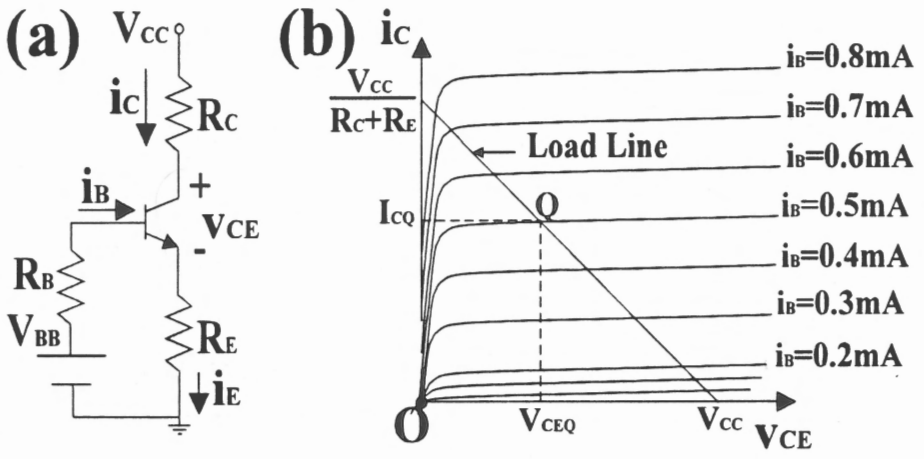


Figure 7.5: Solving a Simple Transistor Circuit

from which we can derive a straight-line relationship between i_C and v_{CE}:

$$i_C = \frac{V_{CC} - v_{CE}}{R_C + R_E} = \frac{-1}{R_C + R_E} v_{CE} + \frac{V_{CC}}{R_C + R_E} \tag{7.5}$$

You plot the load line on the characteristics graph by noting the axes intercepts: when $i_C = 0$, $v_{CE} = V_{CC}$ and when $v_{CE} = 0$, $i_c = V_{CC}/(R_C + R_E)$. The point of intersection of the load line and the curve for $i_B = 0.5$mA is the Q-point (see Figure 7.5b) from which one can read off V_{CEQ} and I_{CQ}, at which the BJT is operating. After that, elementary circuit analysis can get you all the other voltages or currents that you may want. Note that I_{BQ} (which we assumed to be 0.5mA) is actually determined by V_{BB} and R_B, and that when current is flowing from collector to emitter, the BJT is "on" and $V_{BE} \approx 0.7$V.

Chapter 8

Transistors as Switches

8.1 The "On" and "Off" States

As a generalisation, transistors are used in two main rôles in circuits - for either *switching* or *amplification.* In this chapter we consider the transistor as a switch, and a short introduction to transistor amplifiers will be given later in the course.

You already know how the npn transistor can be used as a kind of current tap or switch. If correctly biased (base-emitter forward-biased and base-collector reverse-biased), the transistor will pass current from collector to emitter as the electrons travel in the opposite direction. By controlling the voltage on the base with respect to the emitter, the biasing is altered and the current flow can be cut off. In terms of its emitter-base characteristic, the transistor operates above the knee when "on" and below it when "off". Similarly, if we are referring to its collector-emitter characteristic, the device is in the saturation region when "on" and in the cutoff region when "off".

A transistor being used as a switch should *not* be operated with v_{BE} near the knee voltage, or equivalently in the linear region of its characteristic curve. A switch is a (very) non-linear device, and it will usually be the aim of a designer to drive the transistor either hard-on (i.e. definitely on) or hard-off - one or the other and nothing in-between! Therefore an effective transistor switch will be operated by making v_{BE} a small voltage (zero, or maybe even negative) to cut off current, or making it a voltage *definitely* above 0.7V (say 5V or more, depending on the transistor used) to allow current to flow.

The cut-off voltage, about 0.7V, at which the transistor moves from the

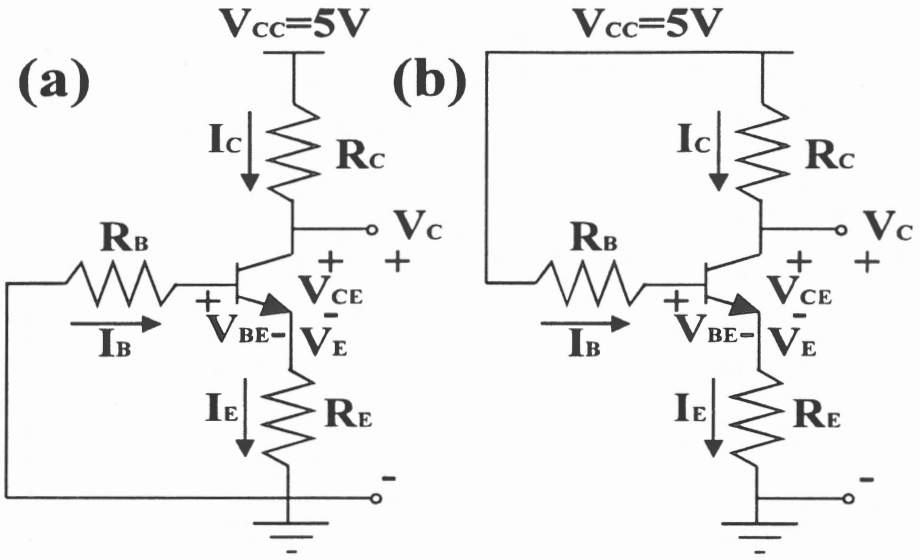


Figure 8.1: (a) The "OFF" State (b) The "ON" State

cut-off to the saturation region, is sometimes written $V_{BE(sat)}$. When the transistor is saturated, the voltage between the collector and the emitter is, of course, small (think of the collector-emitter characteristic), and a typical value at which saturation occurs is $V_{CE(sat)} = 0.2\text{V}$. When in the cut-off region, we have $i_B \approx i_C \approx 0$, in contrast to $i_C \approx \beta i_B$ when the device is "on": note the approximation signs, because there is always a little leakage across junctions, even when the transistor is "off".

It is important to realise that when the transistor is not conducting (i.e. when it is "off"), it effectively presents an open-circuit. This means that the voltage across it, V_{CE}, will be large. In Figure 8.1a, for example, with the base tied to earth, $V_{BE} < 0.7\text{V}$, the BJT is "off", no current flows and so no voltage is dropped across either of the resistors R_C or R_E. Thus $V_C = V_{CC} = 5\text{V}$, $V_E = 0\text{V}$ and $V_{CE} = 5\text{V}$.

Conversely, if the transistor is "on" (see Figure 8.1b), it acts ideally as a short-circuit, and will carry a current given by Ohm's Law as

$$I_C \approx I_E \approx \frac{V_{CC}}{R_C + R_E} \tag{8.1}$$

In the ideal case, $V_{CE} = 0$. More realistically, however, we know that the BJT has dynamic resistance given by

$$r_d = \frac{0.026}{I_E} \tag{8.2}$$

so we will find the V_{CE} has some small value. We can also improve our current calculation for the "on" state to give us

$$I_C = \frac{V_{CC}}{R_C + R_E + r_d} \tag{8.3}$$

In this case, the collector voltage can now be found, again by Ohm's Law:

$$V_C = V_{CC} - R_C I_C \tag{8.4}$$

and this will be a small voltage (because the transistor is "on").

Note. finally, that the voltage at the emitter, V_E, should be used to find the emitter current, and here it helps to realise that the base voltage will be about 0.7V above the emitter voltage while the transistor is "on".

8.2 Switching Larger Currents

We have been looking at examples with $V_{CC} = 5$V supplying the collector of a BJT, but one of the great advantages of the transistor is that small currents and voltages can be used to control much larger ones. There are some transistors which can accept hundreds of volts at their collectors and which can pass current of many amps. These devices would allow, for example, large DC current to be supplied to the load at the touch of a switch in a light-current low-voltage circuit.

Let us try partially analysing the circuit of Figure 8.2, which is meant to represent the way a load requiring 100V (DC) is switched on and off by a small plastic switch that contains an LED. When the switch is put to the "on" position, 1.7V appears across the LED, so $(5.0 - 1.7)/270 = 12.2$mA flows through the switch. With 1.7V on the base of the BJT, its emitter voltage will be about $1.7 - 0.7 = 1.0$V. Then, assuming a small emitter resistance of about 0.1Ω, the emitter current will be $1.0/0.1 = 10$A. The collector current is roughly equal to the emitter current, so we expect about 10A to flow there too. If, now, the switch were to be returned to the "off" position, the transistor's base voltage would be at ground potential, and so

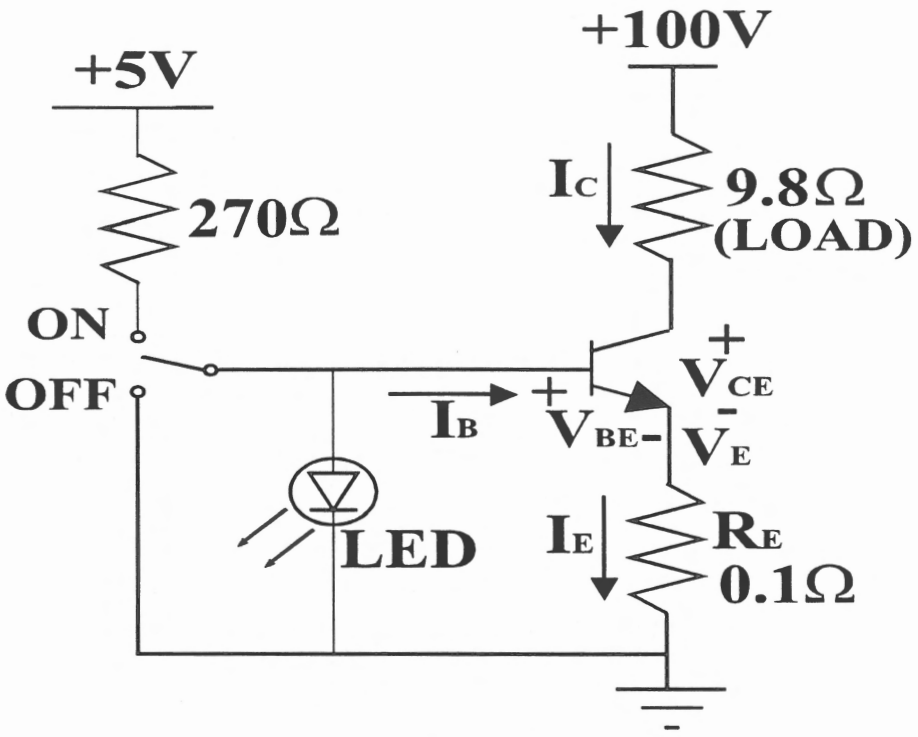


Figure 8.2: A Small Current Switches A Larger One

no current would flow from collector to emitter. Thus, a current of about 10A has been controlled by a mere 5V source and a small switch.

There is a great deal more to this theory of switching, but since it often involves references to the capacitances in the transistor or to the inductance of the load, we shall leave it for later courses.

8.3 Switching and Logic

Instead, let us now turn our attention to transistor *logic*, a most important application of electronic switching. We have already seen how to make the collector voltage go "high" when the transistor is "off", and go "low" when the transistor is "on". In Figure 8.1, "high" and "low" meant about 5V and about 0V respectively. As they denote two very different states for the transistor, we call them *logic levels* (like "true" and "false", or "yes" and "no") and often write them as "1" and "0". You will study the *digital*

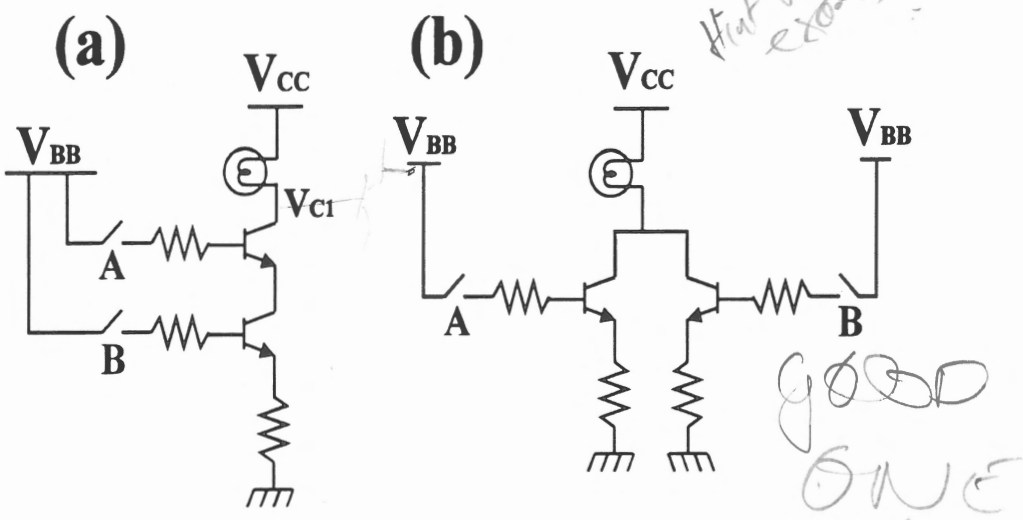


Figure 8.3: AND- and OR-Gate Logical Circuits

and "no") and often write them as "1" and "0". You will study the *digital electronics* that arises from this in some detail in a later course, so we will aim here just to get the flavour of the subject, and to understand the rôle played by the transistor.

Suppose that you wanted a current to flow in a load (e.g. for a light to go on) if two switches (A and B) were "on" simultaneously - in other words, if switch A *and* switch B were both "on". Can you see why the circuit of Figure 8.3a, with transistors in series, can satisfy this need? Only if *both* switches are closed does the voltage V_{C1} go "low", producing the necessary voltage drop across the light, so that current will flow in it.

Similarly, suppose that you wanted a current to make a bell ring if *either* (or *both*) of the two switches was closed. This might be to create a burglar alarm that would be activated if an intruder trod on either of two sensitive pads, each containing a switch A or B. Again, does it seem reasonable that this is achieved, in the circuit of Figure 8.3b, by transistors in parallel? If *either* switch is closed, then a transistor goes "on" and offers a path to ground for the necessary current.

You can build these and similar logic circuits in the laboratory, but when you do so you should recognise that they work as they do simply because they are sets of pn junctions attached (with some resistances) in certain configurations. Since it is merely the type and level of the doping in silicon

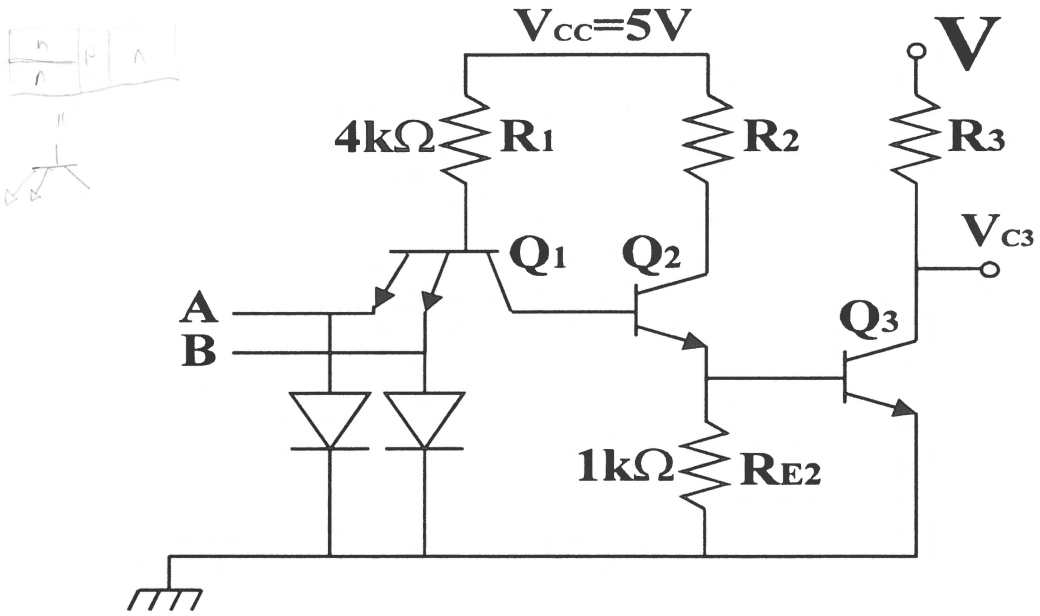


Figure 8.4: A Two-Input Transistor-Transistor-Logic (TTL) NAND Gate

that determines whether a particular piece of silicon is p-type or n-type, or if it is a good conductor (low resistance) or a poor conductor (high resistance), any of these circuits could actually be *integrated* into a single lump of silicon. Inside this single piece, small sections with different doping properties would lie next to each other, giving the same overall effect as the circuit that we have drawn. Such a chip of silicon is, therefore, known as an *integrated circuit* (or *IC*), and this is the form in which logic elements (such as AND, OR, NOT, NAND, NOR gates) are most often to be found. It is worth remembering, however, that the actual circuits are sometimes not much more complicated than the ones in Figure 8.3. We will return to ICs later in the course, when we have discussed the uses of transistors in amplifier circuits.

Before we finish this chapter, try testing your understanding of transistors used as switches in the circuit shown in Figure 8.4. Transistor Q1 is just an npn transistor with its n-type emitter split into two parts, so it effectively has two emitters, as shown.

Now, if either or both of A and B (the inputs) are set to about 0V, then Q_1 has one or two forward-biased base-emitter junctions, and will be "on"

because base current is being supplied to it through the 4kΩ resistor. Its base voltage will be about 0.7V. Current will therefore flow from its collector to the low-voltage emitter(s), and this means that current is drawn away from the base of Q_2. Think about the collector-emitter characteristics of Q_2 and you will see that this must drive Q_2 into its cut-off region. As Q_2 turns "off", its base voltage declines towards 0.7V and Q_1 will become fully saturated (i.e. "on"). With Q_2 "off", V_{BE} for Q_3 is well below 0.7V, so Q_3 is also "off", and its output, V_{C3}, is "high", since no current flows in the resistor R_3: in this context, R_3 is called a *pull-up resistor.*

Conversely, if both A and B are set to V_{CC} (i.e. to 5V), the current flowing into the base of Q_1 will *forward*-bias the base-collector junction, while the base-emitter junction is *reverse*-biased. This is not the way around that transistors are usually supposed to work, but some current will flow towards the base of Q_2. This drives Q_2 into saturation by reverse-biasing its base-collector junction, and Q_2 turns "on" with the voltage across it, $V_{CE2(sat)}$, falling to 0.2V. With current now flowing in the emitter resistor of Q_2 (R_{E2}), the base voltage of Q_3 is raised and Q_3 turns "on". This sends current through R_3, depressing the ouput voltage, V_{C3}.

If you have followed these steps, then you are well on your way to a good understanding of transistors as switches.

Chapter 9

Rectification and Filtering

9.1 Half-Wave Rectification

We now turn our attention back to diodes, to study a few of their most important applications in electronic circuits. We start with *rectification*, which is the process of turning a current or voltage that may be either positive or negative from time to time (an *ac signal*) into one that is always in one direction (a *dc signal*). The diode is particularly suitable for this task because, as we know, it allows current to flow through it in one direction only.

Figure 9.1a shows a circuit containing a 100V sinusoidal ac voltage source. For some of the time, this source inputs a positive voltage into the circuit, and this forward-biases the diode. Assuming for the moment that the diode is *ideal*, it can be replaced, when forward-biased, by a short-circuit. The output voltage, v_o, can now easily be found by voltage division. In the other half of the input cycle, however, the source voltage is negative, so the diode is reverse-biased and acts as an open-circuit. The output voltage must be *zero* at these times, since the current flowing in the load resistor is zero.

Figure 9.1b shows the sinusoidal input graphically, and also indicates what the output voltage *would* have looked like if there had been no diode. By voltage division, it would have been sinusoidal, with an *amplitude* of 90V. In Figure 9.1c, the effect of the diode is shown, and you can clearly see that the output is zero during the periods of reverse-bias. We have achieved what is known as *half-wave rectification.*

Now this output is certainly dc, because it is always in one direction (positive), but it might not be a satisfactory voltage with which to run, for example, a small motor that you wanted to operate at a fairly constant

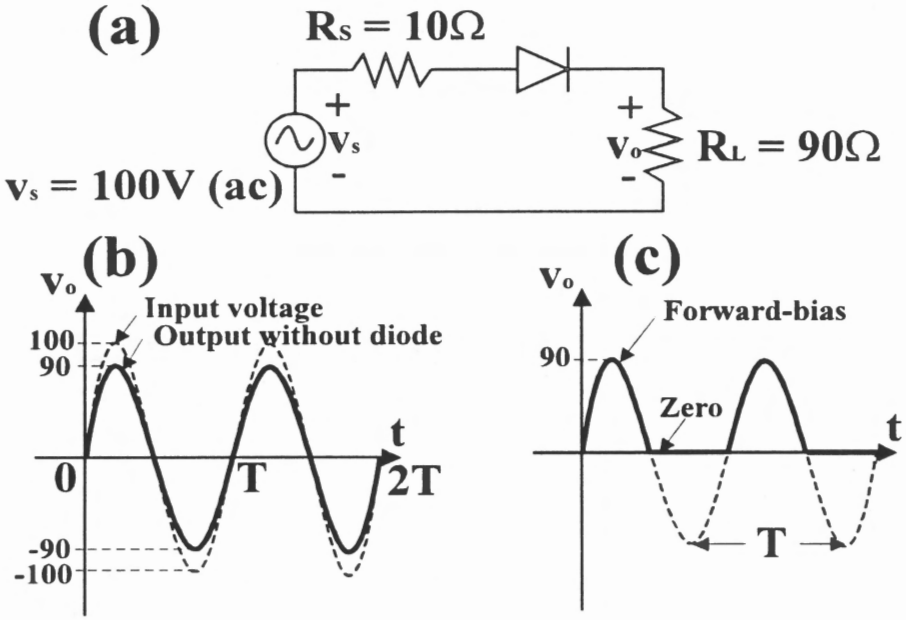


Figure 9.1: Half-Wave Rectification

speed. If you look at the *average* of the input voltage, you should see that it is zero. Our half-wave rectifier has an output voltage whose average is found by adding it up (i.e. integrating it) over one full *period* and then dividing by the period length, T. Thus

$$\begin{aligned} V_{o(avg)} &= \frac{1}{T}\int_0^{T/2} 90\sin\left(\frac{2\pi t}{T}\right)dt \\ &= \frac{-90}{T}\left[\frac{T}{2\pi}\cos\left(\frac{2\pi t}{T}\right)\right]_0^{T/2} \\ &= \frac{-90}{2\pi}(\cos(\pi)-\cos(0)) \\ &= \frac{-90}{2\pi}(-2) = \frac{90}{\pi} \end{aligned} \tag{9.1}$$

This average value is an indicator of the power that is in the half-rectified waveform. The question is, how can we obtain a dc output which has a higher average value than this, which does not fluctuate too much and which does not spend half of the cycle at zero? If we could achieve this, we would have a useful way of converting from ac to dc.

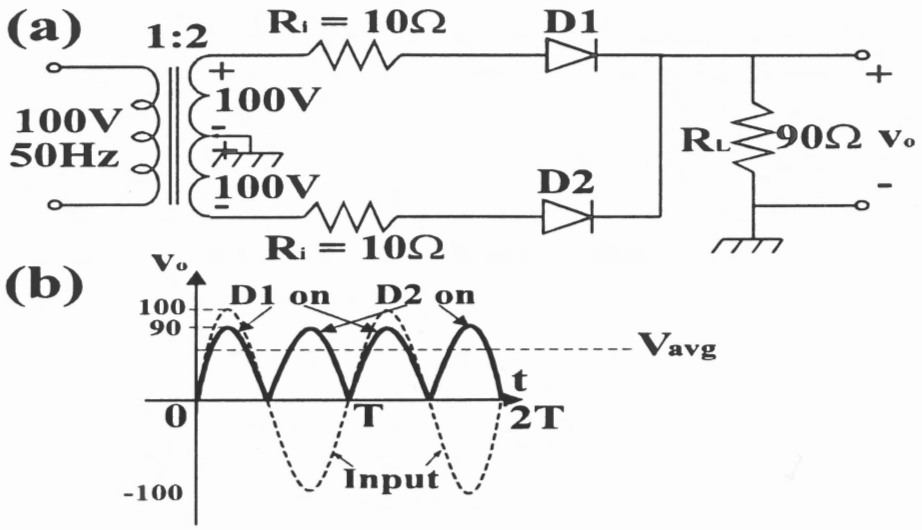


Figure 9.2: Full-Wave Rectification

9.2 Full-Wave Rectification

Part of the solution to this question lies in using a *full-wave rectifier*, which is able to transfer input energy to the output during *both* halves of the input cycle. Figure 9.2a shows one way of building a full-wave rectifier, using two diodes and a *transformer*.

The transformer consists of two large coils placed in close proximity to each other, so that any changing current in the first coil will *induce* a similar current in the second. You will know from physics that, if there are more turns on the second coil, then the voltage induced across it will be greater than the voltage across the primary coil. In this case, the voltage ratio is designed to be 2:1, so if the input ac voltage has an amplitude of 100V, then a 200V sinusoid will appear across the secondary. The secondary has been *centre-tapped*, and its centre point has been earthed, so that the 200V sinusoid is always centred around 0V, appearing as a 100V sinusoid in each half of the secondary.

During one half of the input cycle, diode D_1 is forward-biased and therefore conducts, while D_2 acts as an open circuit. During the other half-cycle, however, D_2 will conduct while D_1 is reverse-biased, so current still reaches the load. The output voltage is shown in Figure 9.2b, and you can see that it is now a dc signal with no half-periods spent at zero. The average voltage

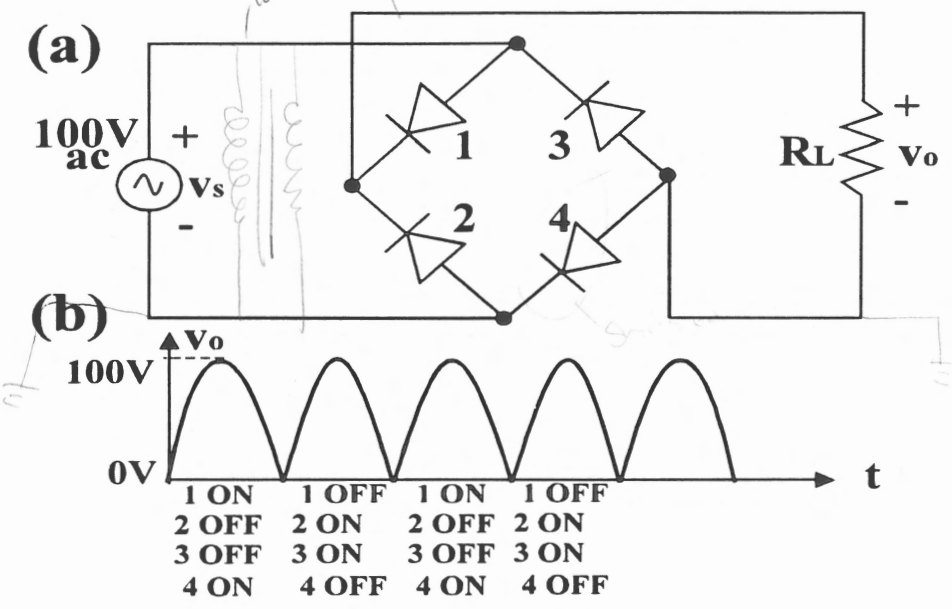


Figure 9.3: The Bridge Rectifier

is twice that of the half-rectified case, as the following working shows:

$$\begin{aligned} V_{o(avg)} &= \frac{1}{T}\left[\int_0^{T/2} 90\sin\left(\frac{2\pi t}{T}\right)\,dt + \int_{T/2}^{T} -90\sin\left(\frac{2\pi t}{T}\right)\,dt\right] \\ &= \frac{90}{T}\left[\frac{-T}{2\pi}\cos\left(\frac{2\pi t}{T}\right)\right]_0^{T/2} + \frac{90}{T}\left[\frac{T}{2\pi}\cos\left(\frac{2\pi t}{T}\right)\right]_{T/2}^{T} \\ &= \frac{90}{2\pi}(-\cos(\pi)+\cos(0)) + \frac{90}{2\pi}(\cos(2\pi)-\cos(\pi)) \\ &= \frac{90}{2\pi}(2+2) = \frac{180}{\pi} \end{aligned} \tag{9.2}$$

Another well-known way to achieve full-wave rectification is shown in Figure 9.3a. This configuration is known as a *bridge rectifier*, and it requires four diodes but, in theory, no transformer. During one half-cycle (while v_s is positive), diodes 1 and 4 are forward-biased while diodes 2 and 3 are reverse-biased. This allows current to flow from the source, through diode 1, the load and diode 4, back to the source. The output voltage is therefore *positive*. In the other half-cycle (v_s is now negative), diodes 1 and 4 are reverse-biased while diodes 2 and 3 conduct. Current now leaves the lower terminal of the

source and travels through diode 2, the load, and diode 3, before returning to the source to complete the circuit. You should see from Figure 9.3 that the output voltage will still be positive.

We seem to have found a way to obtain fully-rectified output of the same amplitude as the 100V input (see Figure 9.3b). In practice, however, you have to remember three things about this circuit:

- The current always travels through *two* diodes whichever way it flows, so the voltage must experience a drop of $2V_\gamma \approx 1.4$V
- There will be some resistance (R_s) associated with the source that will result in a voltage drop. The actual ouput voltage can be calculated, just as in Figure 9.1b, using voltage division.
- The load is not connected to earth in this simple bridge rectifier, and it *cannot* be connected to earth because this would short out one of the diodes. So if the load is to be grounded at all (which may be highly desirable), a transformer may be required to *isolate* the two grounds from each other.

9.3 Filtering

We have now seen that full-wave rectifiers, adequately designed, will convert an ac signal to dc. However, we made the point earlier that we would probably not want to drive many loads with a voltage or current that fluctuates like this. These fluctuations, known as output *ripple*, can be reduced considerably by *filtering.*

Referring to Figure 9.3b, we see an output voltage which varies *twice* each input cycle from 100V to 0V. If this was a rectified waveform from a 50Hz ac signal, this would mean that our output voltage falls from 100V to 0V precisely 100 times per second, taking 1/200 second to do so. Now consider the circuit of Figure 9.4a which is a full-wave rectifier with the resistance associated with the coil, R_i, omitted for simplicity, and with a capacitor, C, connected in parallel with the load. Clearly, this capacitor will charge up as the output voltage rises towards its peak value of 100V. As the output voltage falls again, however, the capacitor will begin to discharge from 100V. It *cannot* do this through the diodes, because they do not permit current to pass backwards. The capacitor must therefore discharge through the load resistor.

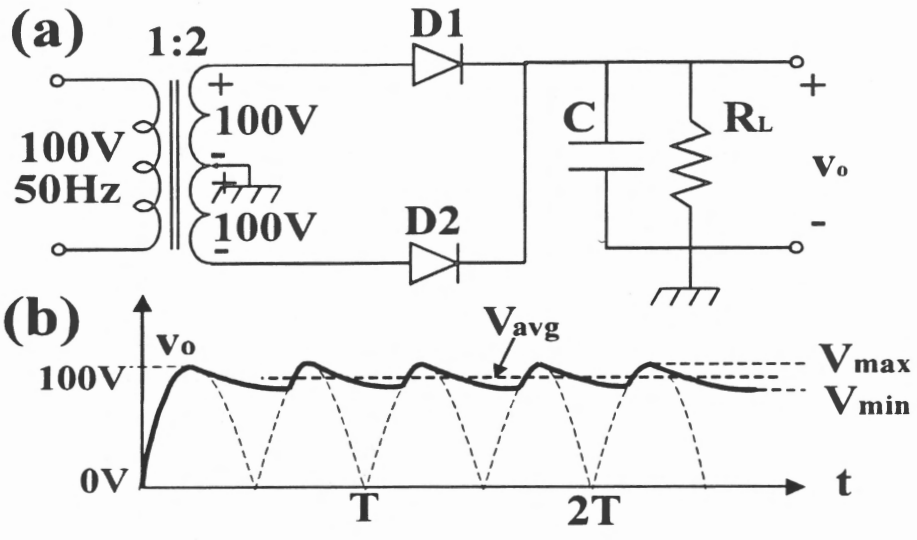


Figure 9.4: Reducing Ripple By Filtering

We know from our recent work in electrical circuits that an RC combination such as this will discharge the capacitor *exponentially*, with the fraction $1/e$ of the charge being left after each *time constant*, τ, has elapsed, where $\tau = RC$. During discharge, the output voltage after time t will be

$$v_o(t) = V_{max}e^{-t/R_LC} = 100e^{-t/R_LC} \tag{9.3}$$

Now, if the capacitance is chosen so that the time constant is a good deal larger than 1/200 second, then the output voltage (which is the same as the capacitor voltage) will fall as shown in Figure 9.4b. As soon as the output from the rectifier rises above the capacitor voltage on its next cycle, the capacitor begins to recharge again, and the output voltage once again rises. We therefore have a way to reduce the ripple. Instead of fluctuating from 0V to 100V, the output now varies from V_{min} to $V_{max} = 100V$, and its average value has been considerably raised (sending more power to the load). The ripple voltage, V_r, can be calculated from the formula

$$v_r = \frac{V_{max} - V_{min}}{2\sqrt{3}} = \frac{\Delta V}{2\sqrt{3}} \tag{9.4}$$

and can be compared to the full dc voltage via the *ripple factor*, where

$$ripple\ factor = \frac{V_r}{V_{avg}} \tag{9.5}$$

How would you go about choosing an actual value for C in a particular application? If you know the load resistance, R_L and the frequency of the input in *hertz*, f, and you require voltage limits V_{min} and V_{max} on your output, then it can be shown that a good safe value for C is

$$C = \frac{V_{max}}{2\Delta V f R_L} \tag{9.6}$$

where $\Delta V = V_{max} - V_{min}$. Alternatively, for a ripple factor certainly below 1%, simply choose C so that the time constant RC is greater than *five* times the period of the input. (Period, T, is $1/f$). Then

$$R_L C > \frac{5}{f} \qquad or \qquad C > \frac{5T}{R_L} \tag{9.7}$$

Chapter 10

Clippers

10.1 Ideal and Practical Diodes

Diodes are frequently used in circuits to eliminate the part of a waveform that lies above or below some *reference* value. Rectification, as we have seen, is a special form of this, where the reference value is the zero level and we wish to eliminate everything below it. The more general case, in which the reference level is arbitrarily set and where one may discard the signal on either side of it, is known as *clipping*. Clipping circuits include a battery and a diode: the battery sets the reference level, and the direction of the diode determines upon which side of the level the signal is clipped off.

Consider the four circuits of Figure 10.1. In each case, we assume that the input, v_i, is a sinusoidal signal of (say) 10V, and that the battery is a 6V dc source. Also, for simplicity, let us begin by assuming that the diodes are *ideal* (we will refine this assumption shortly).

In Figure 10.1a, as the input rises, the diode is (at first) reverse-biased and therefore presents an open circuit. The battery is effectively disconnected so long as the diode is in this "off" state. So, while $v_i < 6$V, no current flows and the output voltage *tracks* the input. As soon as the input rises above 6V, the (ideal) diode becomes forward-biased and now acts as a short circuit. The output is therefore effectively being taken across the battery terminals and so it must equal 6V. This situation holds until the input falls below 6V again and the diode is once again reverse-biased. The output is shown graphically as part of Figure 10.1a, and you can see that it has been *clipped* at 6V.

Clearly, by changing the battery's voltage rating, this threshold could have

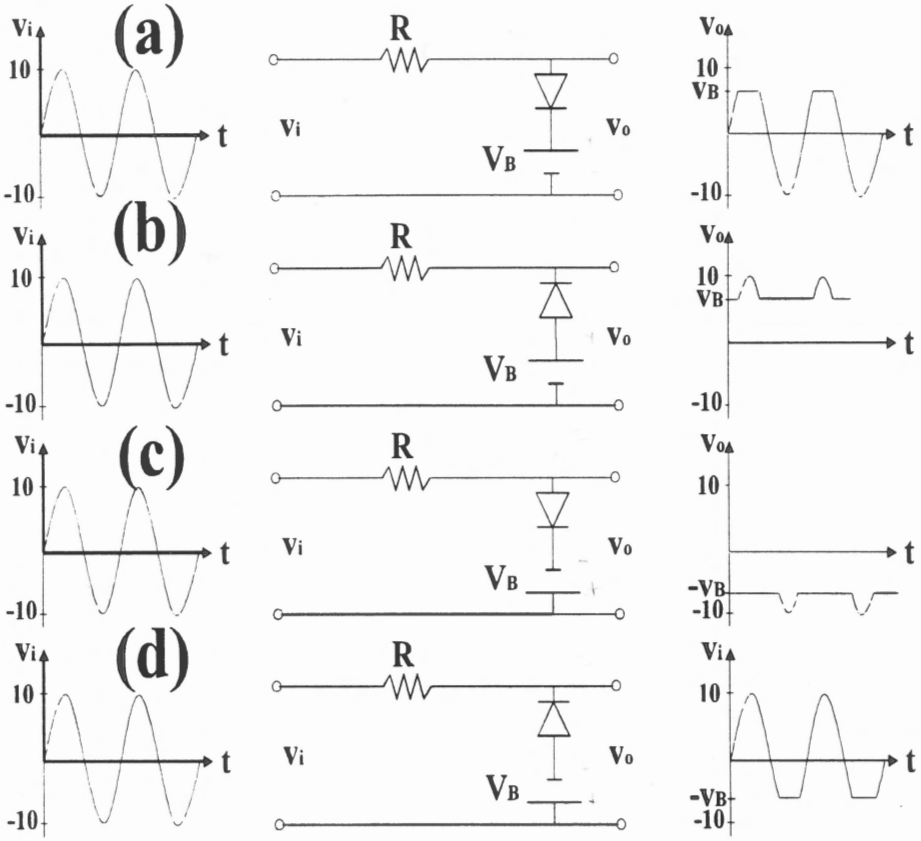


Figure 10.1: Four Ideal Clipping Circuits

been placed anywhere up to the full 10V level, and in Figure 10.1c we see how clipping at -6V is achieved by the simple expedient of turning the battery around. You should examine all four of these circuits carefully, and make sure that you clearly understand how they operate and that you agree with the outputs given in the graphs.

In reality, we know that silicon diodes are not ideal, and exhibit a forward-voltage drop of $V_\gamma = 0.7$V: thus, the diode does not actually switch "on" until the forward voltage is 0.7V, and it maintains this voltage drop while conducting. The effect of this in the circuit of Figure 10.1a would be that the diode would only conduct when $v_i \geq 6.7$V, so the clipping threshold would be *raised*. In Figure 10.1b, however, this consideration would have the effect of *lowering* the clipping threshold to 5.3V. What are the effects of

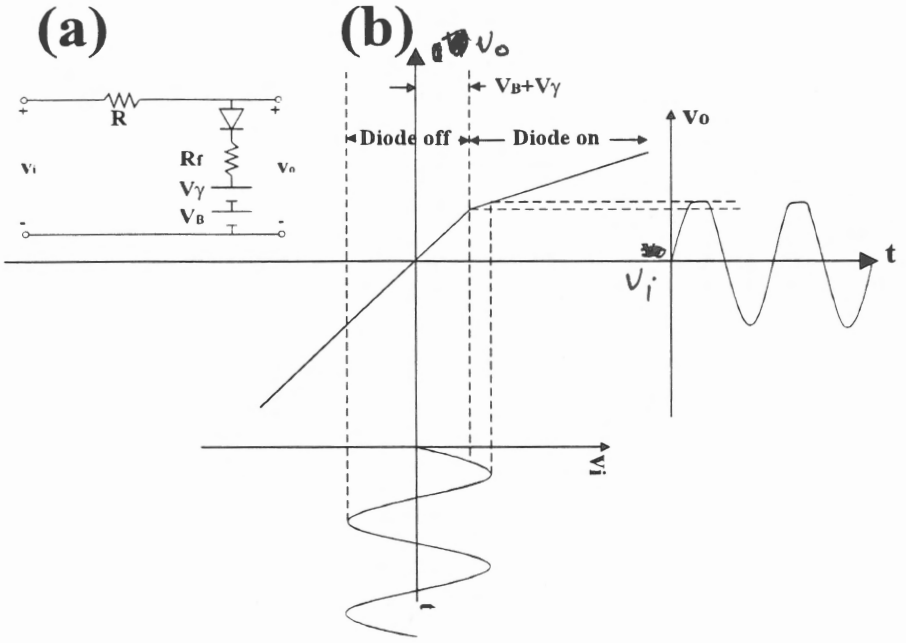


Figure 10.2: A Practical Clipper Model, and the Nature of Clipping

V_γ in the other two cases?

You may recall from our discussion of simple diode models that, when conducting, a practical diode exhibits a forward resistance, R_f. Again, we can use this to refine our expectations of what the clipped output will look like. Because some of the input voltage is dropped across the forward resistance (see Figure 10.2a), the clipped part of the output voltage waveform will not be completely flat. Instead, it will follow the curve of the input, reduced by voltage division in the proportion $R_f/(R_f + R)$. Since R_f is small, this may mean that the clipped section of the output is flat enough to be regarded as horizontal for all practical purposes.

Figure 10.2b shows a section of the *v-i transfer characteristic* of a diode circuit whose diode turns on at $V_{in} = V_B + V_\gamma$, with a corresponding change in the graph's slope from 1 to $R_f/(R_f + R)$. The changing input voltage moves the Q-point up and down the v-i graph, and it should be clear why the change in slope produces a clipping of the output, with the slight curvature being due to the voltage division. If the diode had no forward resistance, the upper section of the graph would be horizontal, and the clipping would be perfectly flat.

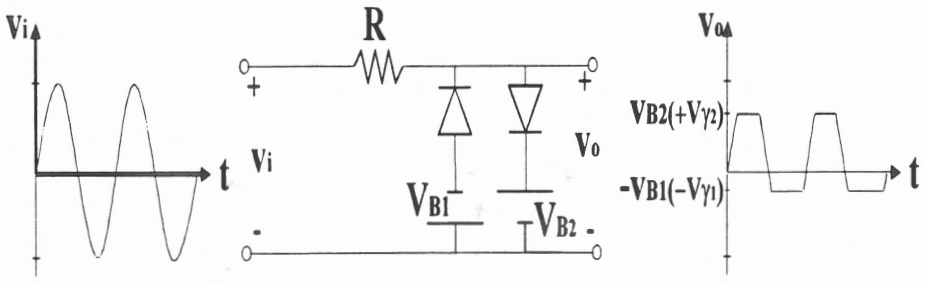


Figure 10.3: A Parallel-Biased Clipper Circuit

10.2 Parallel- and Series-Biased Clippers

Figure 10.3 shows a configuration called a *parallel-biased clipper*, in which positive and negative clipping are performed simultaneously. You should now be in a position to predict exactly what the output would look like, assuming either ideal or practical diodes or, indeed, to be able to *design* a clipper that realises certain desired clipping levels, given an input waveform.

We look finally at the circuits of Figure 10.4, which are four examples of what are known as *series-biased clippers*. With the battery in series with the output, these clipping circuits have the additional effect of raising or lowering the input voltage by some constant (dc) value before clipping it. Consider, for example, the output in the circuit of Figure 10.4a when a 2V ac signal is applied at the input. The diode, if ideal, will conduct only if the voltage above it is below 0V, which means that the input voltage must be below 1V. When the diode conducts, it presents a short circuit, thus tying the output to ground (0V). When the diode does not conduct, the output lies 1V below the input, which gives the clipped output that is graphed in Figure 10.4a.

Once again, you are invited to look over the other series-biased clipper circuits carefully. The outputs shown all assume that the diode is ideal, but you should satisfy yourself as to what the effects of forward voltage drop and forward resistance would be, if they were also to be considered.

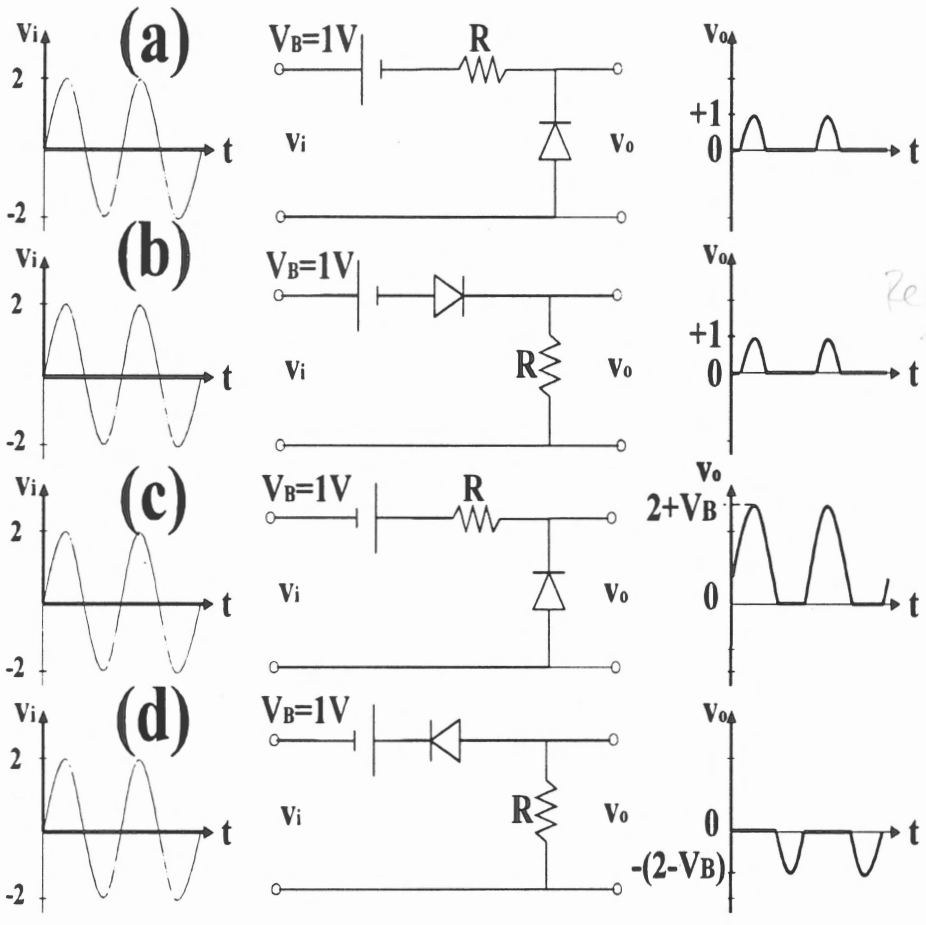


Figure 10.4: Four Ideal Series-Biased Clippers

Chapter 11

Clampers

11.1 Clamping To a Maximum

The final diode application that we shall cover is in a circuit known as a *clamper*. Like the series-biased clipper, this circuit has a shifting effect on the input voltage, raising or lowering it by a certain dc amount. The output of the clamper is not clipped, however, and the amount of the voltage shift is not easy to control.

Instead, with clampers, it is assumed that you have an input voltage that is changing with time and that you would like to preserve its shape. You want, however, to specify that its peak (or perhaps its lowest) voltage should have some particular value of your choice. This is a frequently-encountered need in electronics, because the *shape* of a signal may carry much useful information, and yet the whole signal might need to be brought into a voltage range that conforms to the requirements of the electronic system that is to process it.

Have a look at the clamping circuit shown in Figure 11.1, together with its input and output. The signal has not changed shape (it still has the same amplitude and frequency) but, instead of peaking at V_p, it now goes no higher than the battery voltage, V_B. Let us see how the circuit does this.

It should be clear that, if the output tries to climb above V_B, then the diode (assumed ideal) will begin to conduct, ensuring that the output climbs no higher. While the circuit is in this condition, current flows from the input through the capacitor and diode, and then via the battery V_B back to ground. This has the effect of charging the capacitor, and because the total

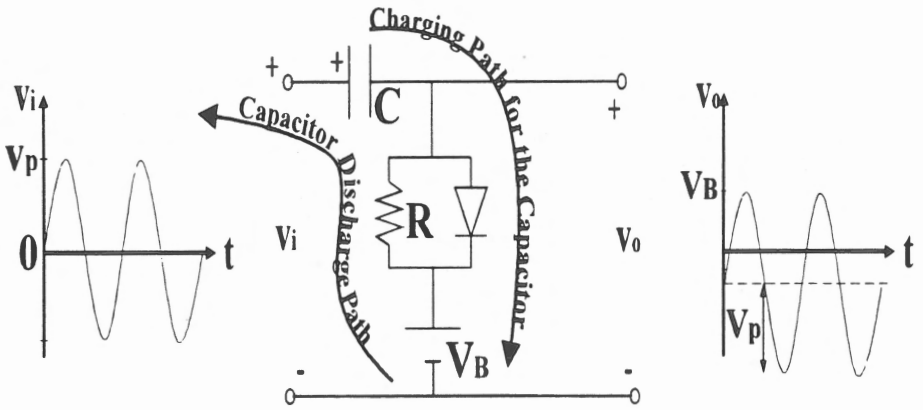


Figure 11.1: Clamping An Input Waveform To V_B

resistance to the current flow is very small (only the internal resistances of the diode and battery plus minute lead resistance) this charging will happen very quickly. By the time that the input reaches its maximum of V_p, the capacitor will have kept up with it and, with V_p on its left and V_B on its right, will be charged to a voltage of $V_p - V_B$.

The input now begins to fall, and the capacitor acts like a battery of voltage $V_p - V_B$. As soon as the input voltage falls, the diode will be reverse-biased and should act as an open circuit. Current will now try to flow from the battery, V_B, up through the resistor and capacitor and back through the source to ground. During this phase, the capacitor will tend to discharge. However, resistor R is intentionally quite large, so the RC time constant of discharge is big, and so discharge is slow. The capacitor can therefore be relied upon to stay as a battery of voltage $V_p - V_B$ for quite a long time. Hence, the output will *track* the input, but at $V_p - V_B$ volts below it.

Whenever the input reaches its peak (even if this is very rarely) it quickly "tops up" the capacitor, with the result that we get a steady dc voltage of $V_p - V_B$ subtracted from the input. The peak of the output is therefore $V_p - (V_p - V_B) = V_B$ volts, and we say that the output peak has been *clamped* to V_B. It is a remarkable feature of this circuit that it adjusts the input waveform without needing to "know" its exact shape in advance. As soon as it "sees" the first input peak, it will clamp that value to the battery voltage, V_B, and output the rest of the waveform in correct relation to it.

Naturally, there are some sources of imperfection which might slightly distort

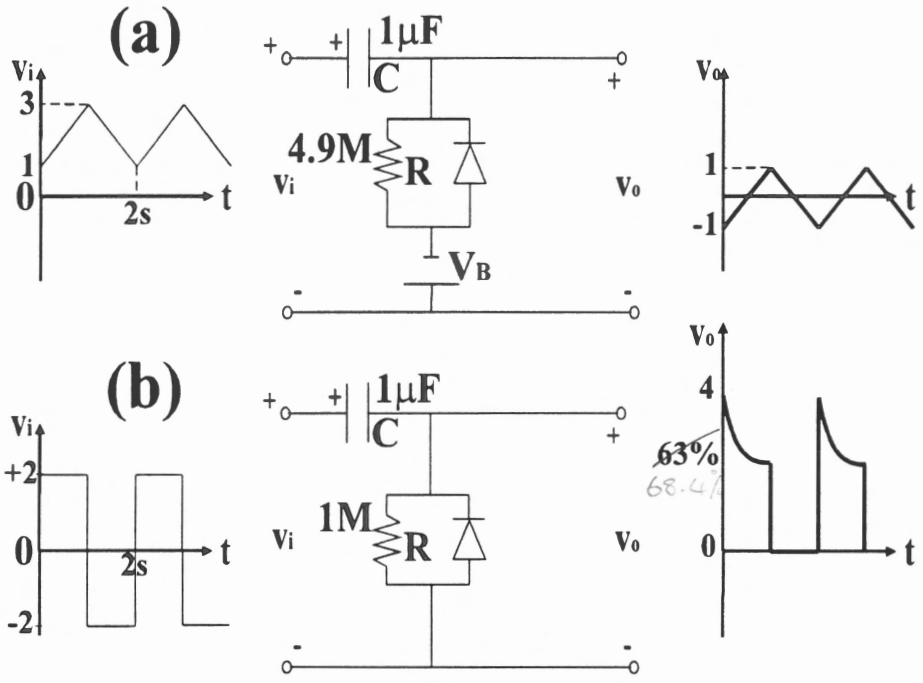


Figure 11.2: Two Circuits That Clamp To a Minimum

the output. These are generally minor, but would include the fact that it takes finite time for the capacitor to charge and non-infinite time for it to discharge. There is also a small leakage current through the diode when reverse-biased that adds to the current discharging the capacitor. As a rule of thumb, however, if the *RC* time constant is kept to *at least* five times the duration of the typical maximum variation in the input (i.e. half a period if it is periodic), then the distortion will be within reasonable limits. You could make *RC* twice this value, just to be sure!

11.2 Clamping To a Minimum

Figure 11.2 shows two more clamping circuits, and in both cases their diodes are in the reverse direction from that of Figure 11.1. This means that their output will clamp to a *minimum* rather than to a maximum value. In Figure 11.2a, a triangular waveform, varying from +1V to +3V, is sent into the clamper circuit, which then clamps its minimum to -1V, giving an output that is symmetrical about the horizontal (0V) axis. Try to identify the

charging and discharging current paths and their directions, and make sure that you know the voltage to which the capacitor will charge, and its polarity. The input triangular waveform has a *period* of two seconds, so we have chosen a 1μF capacitor and a 4.9MΩ resistor. Does this seem reasonable to you?

In Figure 11.2b, a square wave of period two seconds, alternating between +2V and -2V, has its minimum clamped to ground. Regrettably, the designer of the clamping circuit set the *RC* time constant equal to the half-period of the input (1 second). The result is shown in the graph of the output voltage: in half a period the capacitor has time to discharge to ~~$1 - 1/e = 68.2\%$~~ of its peak voltage, with resulting severe distortion of the output waveform. What would you recommend to improve this?

$$\frac{2+\frac{2}{e}}{2+2} \times 100\% = 68.4\%$$

Chapter 12

Transistor Amplifiers (1)

12.1 The Choice of the Q-Point

Transistors used as *switches* are kept in either the saturation ("on") region or the cutoff ("off") region of their characteristic operating curves. If, on the other hand, a transistor is kept in its *linear* region of operation, another extremely useful property of transistor circuits can be exploited - and this is *amplification*. To understand how amplification arises, let us consider the circuit of Figure 12.1a, together with the characteristic curves of the npn transistor which it contains, shown as Figure 12.1b. This particular configuration is known as a *common-emitter* (or *CE*) *amplifier* because, as you may recall from an earlier chapter, the transistor is connected with the output taken from the collector.

The circuit itself simply comprises a transistor with a resistor connected to each of its three terminals. Power is supplied to the collector from V_{CC}, and the base is *biased* by means of the battery, V_{BB}. For the moment, we shall set the ac supply, v_s, to zero. It is a simple matter to add the dc load line to the characteristic curves, starting by applying KVL in the collector-emitter loop:

$$V_{CC} - I_C R_C - V_{CE} - I_E R_E = 0 \tag{12.1}$$

Since $I_C \approx I_E$, this quickly yields

$$I_C = \frac{V_{CC} - V_{CE}}{R_C + R_E} \tag{12.2}$$

which gives the required relationship between the variables i_C and v_{CE}. The intercepts are found by first setting $I_C = 0$, in which case $V_{CE} = V_{CC}$, and

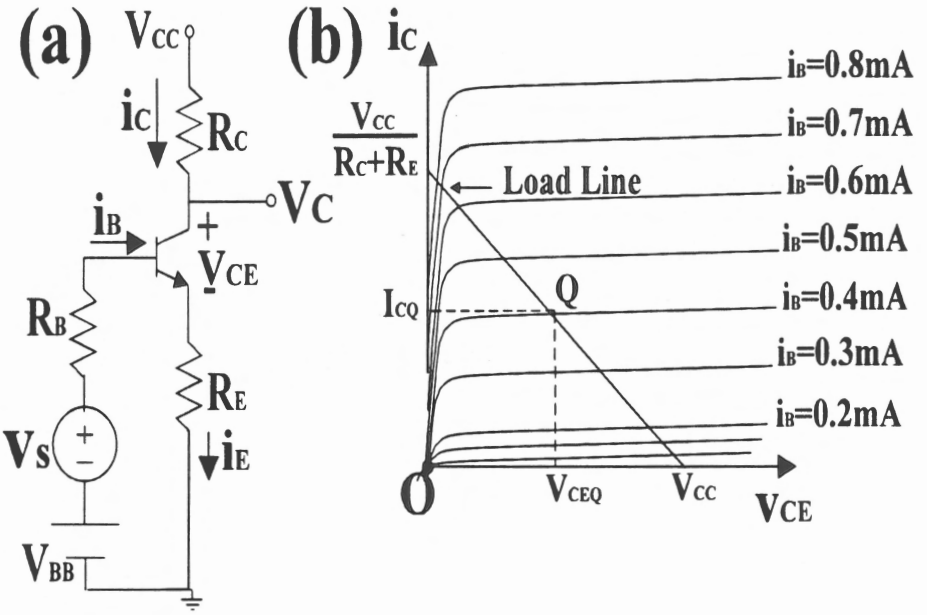


Figure 12.1: A Common-Emitter Amplifier and Its Characteristics Graph

then by setting $V_{CE} = 0$, which would give $I_C = V_{CC}/(R_C + R_E)$. The resulting load line is plotted in Figure 12.1b, and the transistor can be assumed to be operating on a point on this load line that is determined by knowing *one* of I_{CQ}, V_{CEQ} or I_{BQ}.

Now, let us tie down the Q-point more precisely. The whole idea of amplification is that we should allow a small *signal*, v_s, to vary the current into the base, and that this should move the operating point up and down the load line. The result of this movement will be changes in both I_C and V_{CE} which will track the variations in base current, but will be of a much greater amplitude. Evidently, if I_C and V_{CE} are to vary *proportionately* to the changes in I_B, it is important that the transistor be in its linear region, never straying into saturation or cutoff. To maximise the chance that the BJT will stay in its linear range, the Q-point ought to be equally-spaced between the saturation region and the cutoff region at either end of the load line. For this reason, when making a transistor amplifier, we would try to ensure that

$$V_{CEQ} \approx \frac{V_{CC}}{2} \tag{12.3}$$

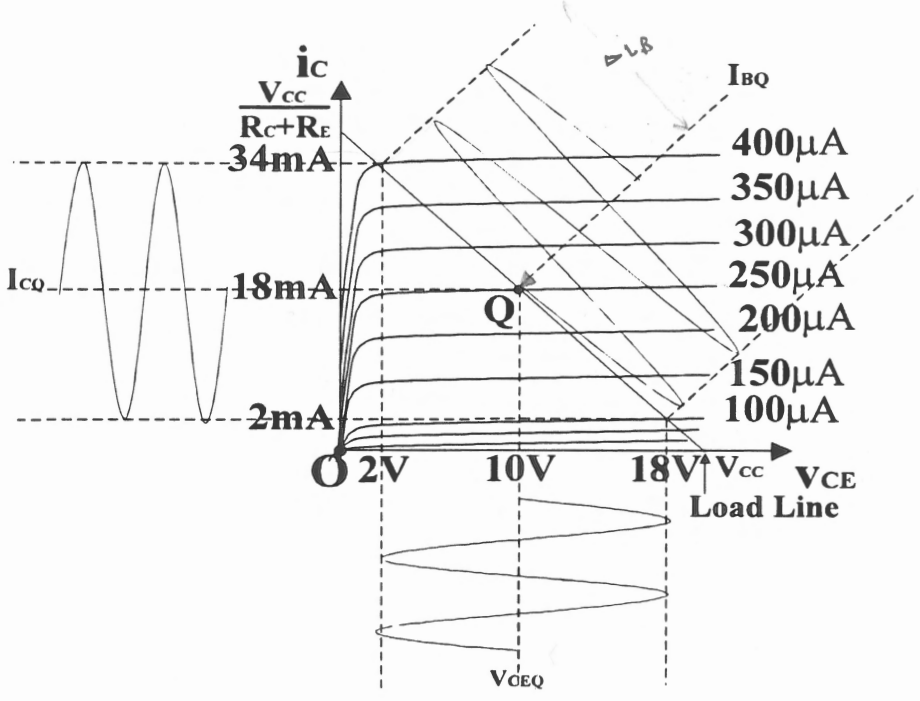


Figure 12.2: Graphical Representation of Amplification

12.2 Amplification of Small Signals

Have a look at Figure 12.2, in which a transistor amplifier has its Q-point chosen at $V_{CEQ} = 10\text{V}$ (as V_{CC} happens to be 20V). Under quiescent conditions, the base current is $I_{BQ} = 250\mu\text{A}$, and the collector current is $I_{CQ} = 18\text{mA}$. We now assume that the input signal, v_s, is a sinusoidal voltage which produces a variation in the base current into the transistor, as shown in Figure 12.2. The amplitude of the base current variation is $\Delta i_B = 150\mu\text{A}$. The operating point will move in response to this variation, travelling sinusoidally up and down the load line, with the Q-point as its average position. The corresponding variations in I_C and V_{CE} are also graphed in Figure 12.2, and you can verify that I_C has amplitude $\Delta i_C = 16\text{mA}$ around I_{CQ} while V_{CE} oscillates about V_{CEQ} with a peak variation of $\Delta v_{CE} = 8\text{V}$.

We have therefore got a circuit into which a current with tiny variations is input, and through whose collector flows a current with identical variations but of much higher amplitude. This is known as *amplification* and it can be

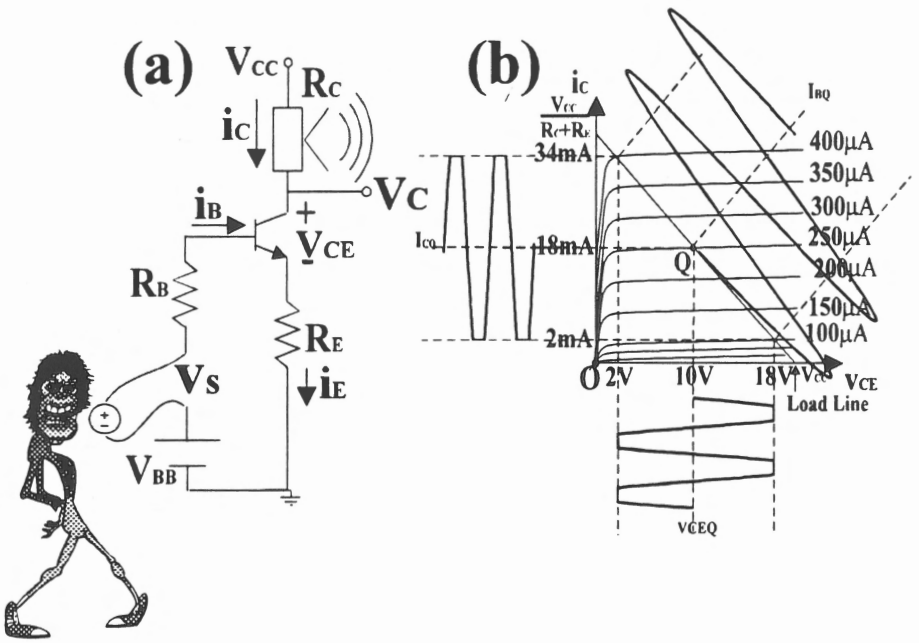


Figure 12.3: Distortion Arises if the Input Amplitude is Too Large

measured by the *current gain*, A_i, which in this case is

$$A_i = \frac{\Delta i_C}{\Delta i_B} = \frac{16mA}{150\mu A} \approx 107 \tag{12.4}$$

Likewise, the *voltage gain* is defined as the ratio of the output voltage swing (Δv_C) over the amplitude of the input voltage fluctuation (Δv_B). In this case, if R_B were a 100Ω resistor, then the voltage swing of v_s required to cause the observed input current swing would be $\Delta v_B = 100\Delta i_B = 15\text{mV}$. The output voltage in the CE amplifier is taken from the collector, and its amplitude is about the same as that of V_{CE} (i.e. $\Delta v_C \approx 8\text{V}$). You will note that this assumes that the emitter resistance is very small - we shall have more to say about that later. The *voltage gain* of the circuit is therefore

$$A_v = \frac{\Delta v_C}{\Delta v_B} = \frac{8}{0.015} \approx 530 \tag{12.5}$$

It is these current and voltage gains which make the circuit (and others like it) so useful to engineers. Consider, for example, if v_s were the small signal

that comes from a microphone - perhaps a signal measured in milliamperes that is quite incapable of driving a loudspeaker. After amplification, however, I_C is a current of over 100 times this size, and a loudspeaker might be placed in the collector circuit, perhaps with R_C representing its resistance. Evidently, we have the basis for a megaphone! (see Figure 12.3a).

It should also be clear that if the input swings are too large, then they will drive the amplifier into saturation (if V_{CE} is too low) or into cutoff (if I_C is too low). Figure 12.3b illustrates the resulting *distortion* of the output. We are all familiar with the distorted sound of an audio system played at high gain (i.e. high volume). Even though this may not be caused by clipping in the amplifier, it gives you the impression of why this type of signal distortion is undesirable!

12.3 Some Refinements to the CE Amplifier

Referring again to Figure 12.1a, let us look in more detail at some of the relationships between the quantities in the common emitter amplifier. We can apply KVL in the base loop to get

$$V_{BB} - I_B R_B - V_{BE} - I_E R_E = 0 \tag{12.6}$$

If we note that $I_E \approx I_C$, that $I_C = \beta I_B$ in the linear range, and that $V_{BE} = V_\gamma \approx 0.7\text{V}$ for silicon, we see that

$$V_{BB} = \frac{R_B I_C}{\beta} + V_{BE} + I_C R_E \tag{12.7}$$

Thus, at the Q-point, we can re-write this as

$$I_{CQ} = \frac{V_{BB} - V_{BE}}{R_B/\beta + R_E} \tag{12.8}$$

Now, there are three refinements to the basic amplifier of Figure 12.1a, which need to be well-understood. The first of these is incorporated into the circuit of Figure 12.4a, in which the dc input voltage, V_{BB}, has been replaced by two resistors so as to supply the BJT base with a fraction of V_{CC} by voltage division. Evidently, this has the advantage of saving a voltage source and, properly designed, it is a good substitute for V_{BB}. The variable (ac) signal, v_s, is *superimposed* on the *dc bias voltage* due to these resistors, as seen in the diagram.

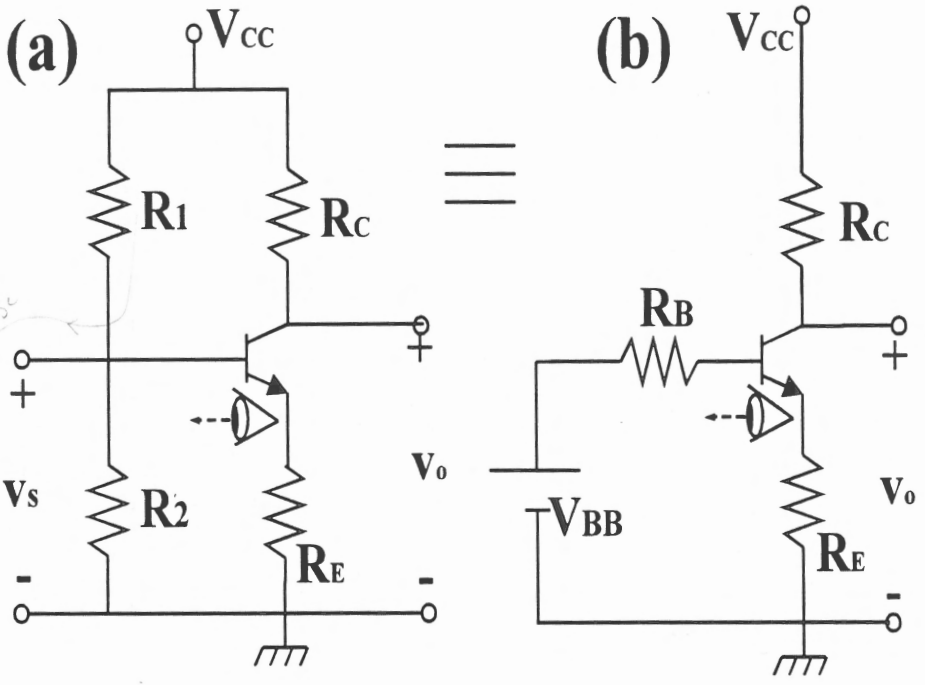


Figure 12.4: Voltage Divider Biasing, And Its Thévenin Equivalent

You should be clear that, looking back from the base of the transistor, the Thévenin equivalent "seen" *by an ac signal* is as shown in Figure 12.4b. You start by setting all ac sources to zero, and the Thévenin voltage is then the open circuit voltage seen at the base, namely

$$V_{TH} = V_{OC} = V_{BB} = \frac{R_2 V_{CC}}{R_1 + R_2} \tag{12.9}$$

Then, because the short-circuit current is simply

$$I_{SC} = \frac{V_{CC}}{R_1} \tag{12.10}$$

you can obtain the Thévenin resistance as

$$R_{TH} = \frac{V_{OC}}{I_{SC}} = R_1 // R_2 = R_B = \frac{R_1 R_2}{R_1 + R_2} \tag{12.11}$$

These important relationships, argued from the Thévenin equivalent circuit, can be used to give equations for the biasing resistors R_1 and R_2, assuming

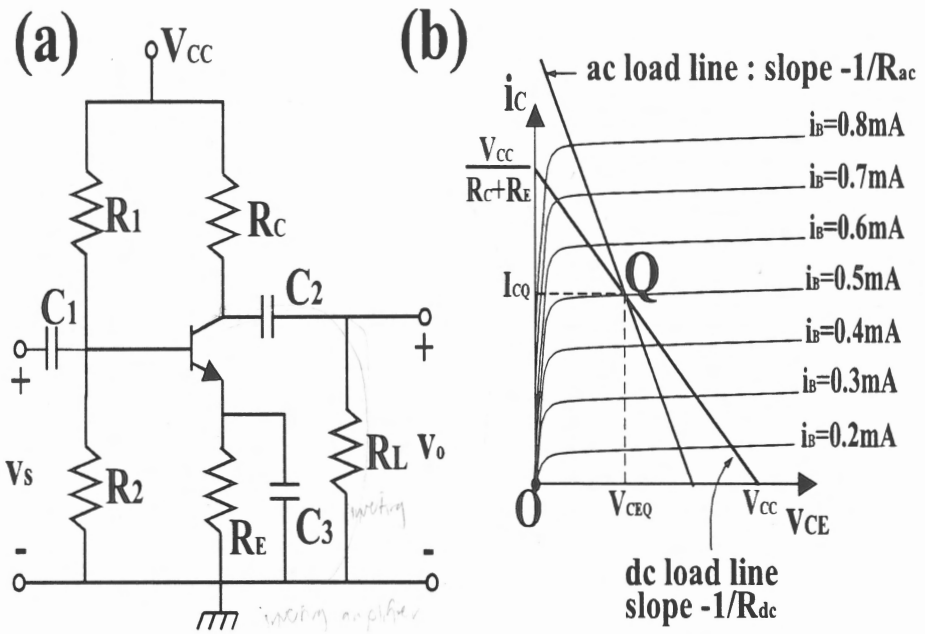


Figure 12.5: A CE Amplifier and Its AC and DC Load Lines

that you know what values of V_{BB} and R_B gave you a suitably-positioned Q-point. You should check that you agree that the same Q-point is achieved if

$$R_2 = \frac{R_B V_{CC}}{V_{CC} - V_{BB}} = \frac{R_B}{1 - V_{BB}/V_{CC}} \tag{12.12}$$

and

$$R_1 = \frac{V_{CC} R_B}{V_{BB}} \tag{12.13}$$

Our second refinement to the basic amplifier is shown in Figure 12.5a. As you can see, three capacitors have been added to the amplifier circuit, and the output is now taken over a load resistor, R_L. Capacitors C_1 and C_2 are called *coupling capacitors*. As we know, capacitors tend to pass ac signals but present an open circuit to dc. C_1 and C_2 are large capacitors, intended to allow the ac input and output signals to go through the amplifier and not to permit any dc from externally-attached circuitry to affect the important dc conditions on the transistor's base and collector (which we know are what make the amplifier work at all, by keeping the BJT in its linear region).

Capacitor C_3 is called a *bypass capacitor*, and its rôle is to let ac signals at the emitter bypass the emitter resistor (ac will prefer to travel through

the capacitor than to go through R_E, but dc will take the resistor route because the capacitor looks like an open circuit to it). The result of this is that the amplitude of any voltage variation at the collector now *exactly* equals the variation in V_{CE}, because none of the voltage gain is lost in R_E. In other words, the bypass capacitor makes sure that the ac voltage gain of the amplifier is as big as possible.

The final refinement that needs to be made to our theory concerns the load line of the amplifier, along which the operating point moves in response to the input signal. We have become familiar with obtaining the *dc load lines* for diodes and transistors, which is really the line of all possible Q-points for all possible bias situations. The input signal, however, is generally an *ac signal*, and if the amplifier contains any capacitors, then the load for an ac signal may well be different to that for a dc signal. Take the circuit of Figure 12.5a, for example. Hopefully you agree that its dc load line will have a slope given by

$$\frac{-1}{R_{DC}} = \frac{-1}{R_C + R_E} \tag{12.14}$$

and you will have no trouble in plotting it on a graph. By contrast, ac signals emerging at the collector "see" routes to ground (with dc sources short-circuited) via either R_C or R_L. The capacitor across R_E shorts out this third possible ac route. Hence, the resistance down to ground seen by any ac signal is

$$R_{AC} = R_C // R_L \tag{12.15}$$

The corresponding *ac load line*, therefore, has a slope of

$$\frac{-1}{R_{AC}} = -\left(\frac{R_C + R_L}{R_C R_L}\right) \tag{12.16}$$

The ac load line must pass through the Q-point, so it is quite easy to construct it on a graph, knowing also its slope. At this stage, you should just satisfy yourself that you could do so (given R_C and R_L) and that the ac line is simply a more accurate (i.e. realistic) load line to be using where ac signals are concerned than our original dc load line. Figure 12.5b shows both load lines plotted on a characteristics graph for a typical amplifier.

Chapter 13

Transistor Amplifiers (2)

13.1 An Analysis Method For CE Amplifiers

We are now in a position to discuss a formal method for *analysing* a CE amplifier. You ought to be aware that *analysis* means taking a given circuit, in which you know all the component values, and then using your understanding of how it works to predict various currents, voltages, gain factors etc. Analysis is different from *design*, which means building the circuit from scratch in order to meet certain stated requirements (e.g. to achieve a given gain). Of the two, design is the more difficult, so we will leave it for future courses. Our aim here is just to develop a very simple approach for analysing a CE amplifier. You are advised to be sure to understand the full derivation of all of the equations in this analysis, however, because intelligent analysis does *not* mean taking a kind of "recipe-book" approach to a circuit. The idea is that you should use the following method as an example of what to do if confronted with a *similar* (but not identical) circuit.

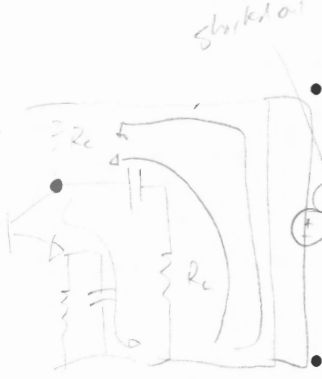
Suppose, for example, that you wanted to analyse the circuit of Figure 12.5a, in the previous chapter and that you knew (as you generally will) the values of V_{CC}, V_{BE}, R_1, R_2, R_E, R_C, R_L, and β. You would proceed as follows:

- **Step 1.** Use R_1 and R_2 to determine V_{BB} and R_B, using

$$V_{BB} = \frac{R_2 V_{CC}}{R_1 + R_2} \qquad \textit{and} \qquad R_B = R_1 // R_2 = \frac{R_1 R_2}{R_1 + R_2} \tag{13.1}$$

- **Step 2.** Now find the quiescent value of I_C (called I_{CQ}), using

$$I_{CQ} = \frac{V_{BB} - V_{BE}}{R_B/\beta + R_E} \tag{13.2}$$



(Remember that $V_{BE} \approx 0.7\text{V}$ for silicon)

- **Step 3.** Obtain V_{CEQ}. This you can do either by drawing the dc load line and reading off V_{CEQ} from the known value of I_{CQ} or by writing the KVL equation for the collector-emitter loop:

$$V_{CC} - I_{CQ}(R_C + R_E) - V_{CEQ} = 0 \qquad so \qquad V_{CEQ} = V_{CC} - R_{DC} I_{CQ} \tag{13.3}$$

  Draw the dc load line now, if you have not already done so.

- **Step 4.** Knowing the Q-point, calculate R_{AC} by finding all the paths that an ac signal could take from the collector to ground. Construct the ac load line on your graph, as the line of gradient $-1/R_{AC}$ that passes through Q.

- **Step 5.** Look at your graph and find the greatest possible swings in I_C, V_{CE} and I_B. Remember that you want to avoid the operating point leaving the linear region. Use this information to find either the current gain or the voltage gain of the amplifier. Remember that the output current and voltage amplitudes are related by $V_C = I_C R_{AC}$.

You are now able to look at the circuit of a CE amplifier and to predict the gain that it will give, the output voltage or current, and where its operating point lies. You should review the last chapter carefully, because it is rather full of new concepts, and try to satisfy yourself that you have understood *why* these amplifiers perform as they do. Then look for examples to analyse, so as to get practice at the use of the equations developed here, and of the load-line technique.

13.2 The Common-Collector Amplifier

Figure 13.1 shows an npn transistor connected as part of a different type of amplifier from the common-emitter circuits we have been studying. In this circuit, the output is taken from the emitter rather than from the collector, and so the configuration is called a *common-collector* (or CC) *amplifier*. You will also very often see it described as an *emitter follower*, because the voltage at the emitter will usually *track* any voltage changes at the base, except that it remains about 0.7V below the base voltage, so long as the BJT is "on". For this reason, the voltage gain of the CC amplifier is very close to *unity* (in fact slightly less, if we consider the voltage dropped across

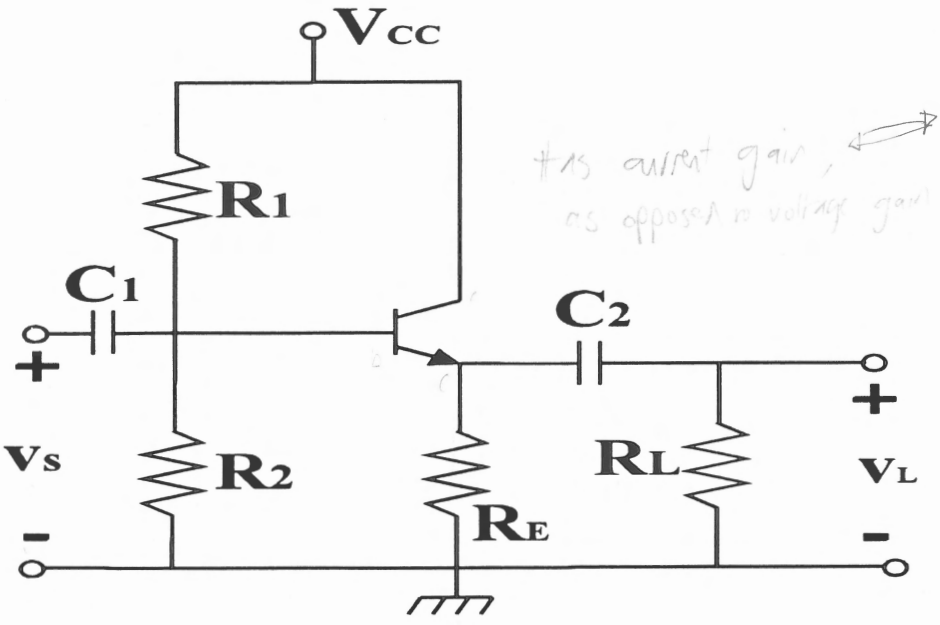


Figure 13.1: A Common-Collector Amplifier or Emitter Follower

the forward resistance of the base-emitter junction within the transistor). You may well wonder what the point of such a circuit is, because it seems to have an amplification factor that actually represents a loss!

The CC amplifier does have some useful features. Firstly, as the input rises so does the output, and they also fall together. Had you noticed that, by contrast, the CE amplifier *inverts* the input signal at the voltage output? Look at any of the load-line graphs of the previous chapter and you can verify that, as the input, V_B (or I_B), increases, V_{CE} falls and thus *lowers* the collector voltage that we regard as the output. Likewise, if the base voltage (or current) falls, then the output rises. You may see this described as a *180° phase shift.* Hence the CE amplifier is an *inverting amplifier*, while the CC amplifier produces no phase shift and is *non-inverting.*

Secondly, the CC amplifier is useful in providing *current gain*, even though its voltage gain is poor. As long as the BJT is "on", it can supply substantial current flowing from collector to emitter, in response to a small change in base current. Note that the dc component of this current "sees" only R_E (there is *no* R_C), while the ac component, which can pass through the coupling capacitor at the emitter output, "sees" R_E and R_L in parallel.

In other words,

$$R_{DC} = R_E \qquad and \qquad R_{AC} = R_E//R_L \tag{13.4}$$

Hence, if R_E is small, and since a dc current of V_{CC}/R_E will flow if the transistor is on, the output current can indeed be substantial.

A third value of CC amplifiers, arising from this, is their ability to give good *power gain.* As the name implies, power gain is simply the ratio of the power supplied at the output of the amplifier to the power at the input. Now, the instantaneous output power is found by looking at the voltage across and the current through the load

$$P_{out} = V_L I_L = \frac{V_L^2}{R_L} = I_L^2 R_L \tag{13.5}$$

while the instantaneous input power is found by examining the voltage and current supplied by the source

$$P_{in} = V_s I_s \tag{13.6}$$

Since, as we have seen, $V_L \approx V_s$, the high current gain of the emitter follower configuration leads directly to its high power gain, A_p, because

$$A_p = \frac{V_L I_L}{V_s I_s} \approx \frac{I_L}{I_s} \tag{13.7}$$

You should not confuse either input or output power with the power dissipated by the transistor itself. This, as you might expect, has instantaneous value

$$P_{BJT} = V_{CE} I_C \tag{13.8}$$

and, in particular, the *average power dissipation* in the BJT is found by examining the Q-point:

$$P_{BJT(avg)} = V_{CEQ} I_{CQ} \tag{13.9}$$

These concepts of power apply equally to CE amplifiers as to CC amplifiers, but you have to be careful that you know where the output is being taken from in each case. The analysis of emitter follower circuits also takes a very similar form to that of the CE amplifier given in the previous section. In fact, the only difference in the two analyses is the values used for R_{AC} and

R_{DC}. Thus, the dc load line of the emitter follower is found, as you might expect, from the equation

$$I_C = \frac{V_{CC} - V_{CE}}{R_{DC}} \tag{13.10}$$

while for maximum swing the Q-point of the CC amplifier should be at

$$I_{CQ} = \frac{V_{CC}}{R_{AC} + R_{DC}} = \frac{V_{CC}}{R_E//R_L + R_E} \tag{13.11}$$

and

$$V_{CEQ} = I_{CQ}R_{AC} = I_{CQ}(R_E//R_L) \tag{13.12}$$

Chapter 14

PNPBJTs, JFETs and MOSFETs

14.1 The pnp BJT

Up until now, all of our examples and explanations have been in terms of the npn transistor, which we have seen in use in both switching and amplification circuits. This is an appropriate moment to mention that there are other forms of transistors, capable of fulfilling the same functions, but which operate by different principles from the npn BJT. Since these other devices are common, and indeed sometimes have distinct advantages, you should at least be familiar with them, and that is the aim of this chapter.

We begin with the pnp transistor, which was mentioned when we introduced transistors, but which we have since deliberately ignored to avoid confusion. The pnp device (sketched in Figure 14.1a together with its circuit symbol) consists of a heavily-doped medium-sized p-type emitter, a thin medium-doped n-type base and a lightly-doped, large p-type collector. The most common material for fabricating these transistors, and all the others we see in this chapter, is silicon, the special qualities of which, when correctly doped, result in the potential hill diagram shown dotted in Figure 14.1b.

When a negative voltage, V_{CC}, supplies the collector, and the base is supplied with a smaller negative voltage so that the base-emitter junction is *forward-biased* and the base-collector junction *reverse-biased*, the transistor switches "on" and current will flow. This current is best visualised as a flow of holes injected from the emitter into the base, climbing a potential slope to do so, and then falling down a second potential hill to the collector. By

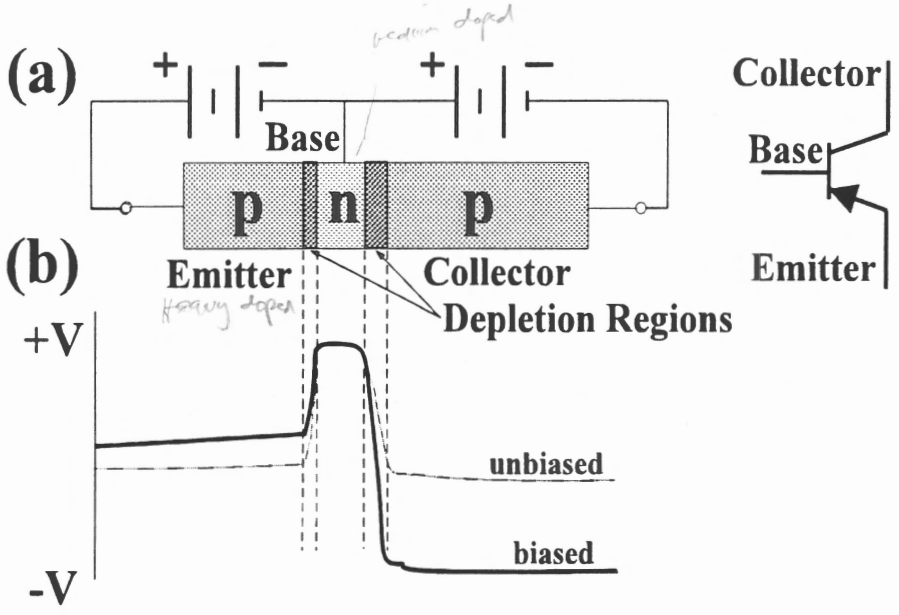


Figure 14.1: The pnp Transistor and Its Potential Hill

varying the base voltage, the height of the first potential hill can be controlled, and in this way current may either flow or be cut off. Figure 14.2a gives the emitter-base characteristic: in a silicon pnp transistor, the knee voltage is typically at $V_{BE} \approx -0.7\text{V}$.

One gets used to the policy of "reversing everything" when thinking about the pnp BJT in a given situation. It ought to be clear from Figure 14.2a that the pnp transistor can be used as a switch, provided that it is driven hard into saturation or cutoff by $V_{BE} << -0.7\text{V}$ or $V_{BE} >> -0.7\text{V}$ respectively. Likewise, the collector-emitter characteristics are just a mirrored version of the npn case, and are shown in Figure 14.2b. You will note that V_{CC}, supplied to the collector, is a *negative* voltage. This means that in an amplifier containing a pnp transistor (such as the simple one shown in Figure 14.2c) current flows upwards from the emitter to the collector, as suggested by the arrow in the BJT symbol. The load line for such a circuit is easy to find. KVL gives

$$-I_E R_E + V_{CE} - I_C R_C = V_{CC} \tag{14.1}$$

so, since $I_E \approx I_C$, we get

$$I_C = \frac{V_{CE} - V_{CC}}{R_C + R_E} \tag{14.2}$$

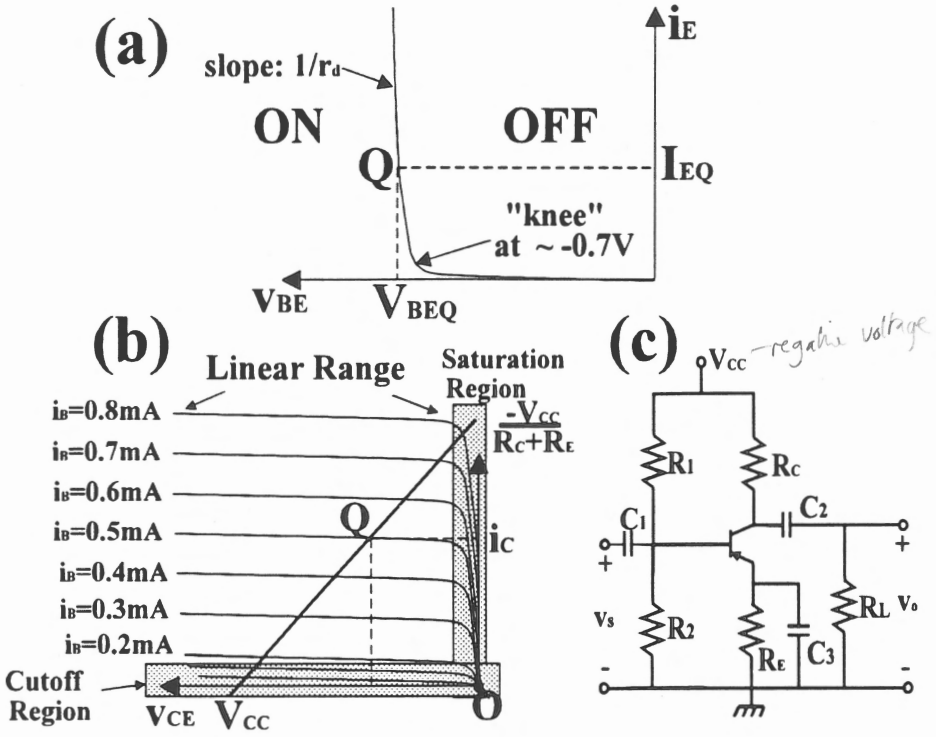


Figure 14.2: Switching and Amplification Characteristics of the pnp BJT

Then, if $V_{CE} = 0$, $I_C = -V_{CC}/(R_C + R_E)$ (which is a *positive* number), while if $I_C = 0$, $V_{CE} = V_{CC}$ (a *negative* number). The load line is plotted in Figure 14.2b, and it should be clear that the pnp-embedded circuit has all of the amplification properties of its npn counterpart, provided that it is correctly biased in its linear region.

Analysis of pnp-embedded amplifiers should follow naturally from this, and you ought to be able to add the pnp transistor to your repertoire of familiar devices relatively painlessly, after careful consideration of *why* it works like it does, and with a little practice! In later courses, you will frequently see the pnp BJT in more sophisticated amplifier and switch designs, so it will repay you to get to grips with its basic operation now.

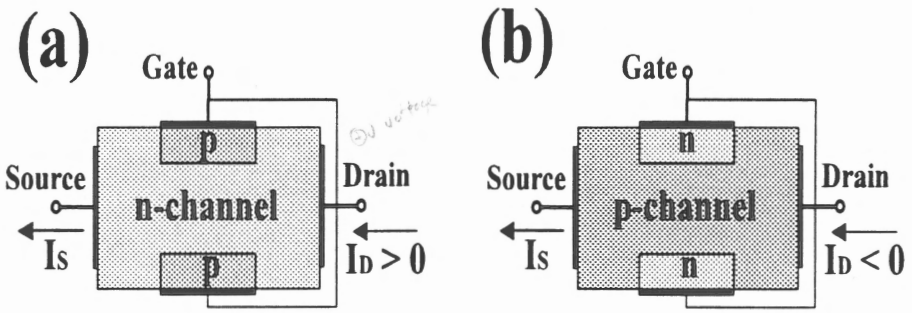


Figure 14.3: The n-channel and p-channel JFETs

14.2 The JFET

The switching and amplification functions of the BJT can also be performed by another class of devices, known as *field-effect transistors* or *FETs*. The first of these, the *junction FET* or *JFET*, is illustrated in Figure 14.3. As you can see, it consists of a strip of doped silicon, known as the *channel*, with two pieces of the opposite doping diffused into it to form the *gate*. Connections are made to the two ends of the channel: a supply voltage, V_{DD} is connected to the end known as the *drain*, while the other end, or *source*, is often connected to ground. Figure 14.3a illustrates the *n-channel JFET*, while Figure 14.3b shows the *p-channel* version of the device.

Now V_{DD} provides a drain-source voltage, V_{DS}, that causes a drain current, I_D, to flow from drain to source through the channel, and to emerge as the source current, I_S. Because the channel has some resistance, this current flow causes a voltage drop between the drain-gate junction and the gate-source junction. This reverse-biases the drain-gate pn junction and thereby produces a *depletion region* around the drain-gate end of the device, shaped as shown in Figure 14.4a. The existence of the depletion region means that the carrier-rich channel is somewhat thinner, and so the channel has a higher *resistance* (look at the equations at the end of Chapter 1).

We can draw a graph of the effect on the drain current, I_D, of increasing the drain-source voltage, V_{DS}. As you see in Figure 14.4b, as V_{DS} is increased from zero, I_D increases linearly at first, as though the resistance of the channel were constant. Then a point is reached where the depletion region starts to "pinch off" the channel, and the increased channel resistance is seen as a decrease in the slope of our graph. Finally, total *pinch-off* of the channel occurs, and after that it doesn't matter if V_{DS} is further increased - the

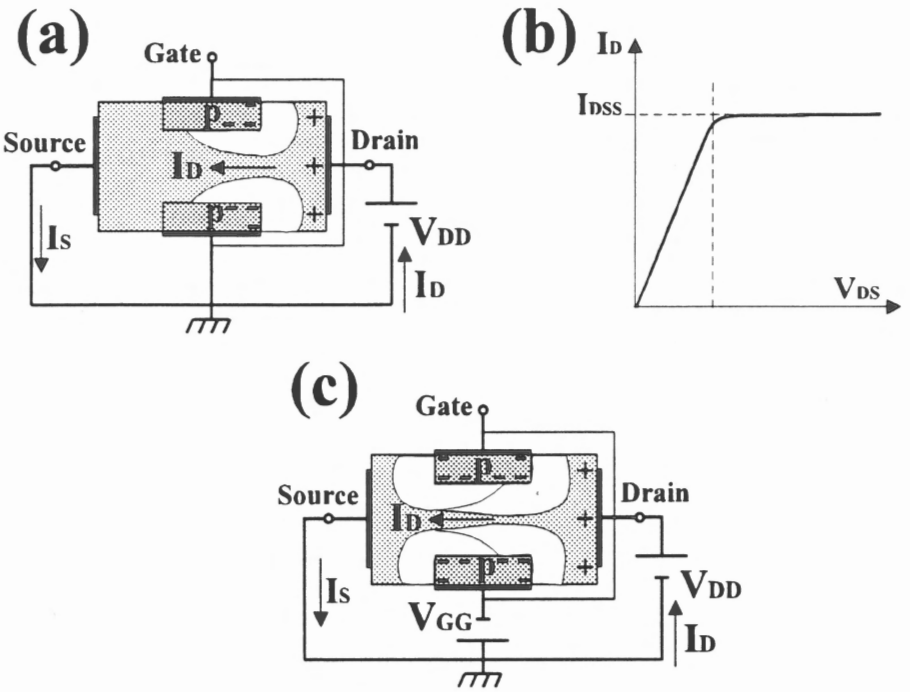


Figure 14.4: The Phenomenon of Pinch-Off in a JFET

channel is said to be *saturated* and a steady current called the *drain-source saturation current*, I_{DSS}, will flow regardless of V_{DS}. (Actually, there comes a point where V_{DS} is so high that the JFET suffers *avalanche breakdown*, but we will not consider that condition because it is of no immediate practical use).

You will immediately notice that we have a curve that looks like a BJT collector-emitter characteristic, and this suggests that the JFET may be used in the same ways as the BJT. If we now include a gate-source biasing voltage, V_{GS}, we get a further way to control the JFET's behaviour. If V_{GS} is ever made negative, this will reverse-bias the gate-source junction (of an n-channel JFET) and produce a depletion region at the gate-source end of the device (see Figure 14.4c). Any applied drain-source voltage will now lead to pinch-off of the channel at a lower value of drain current, I_D, than before, because it is the flow of I_D that causes the voltage drop that causes the depletion region to grow. Conversely, if V_{GS} is increased again, then the depletion region will be reduced by the reaction at the gate-source end, so that more drain current will be required before an effect is seen at the

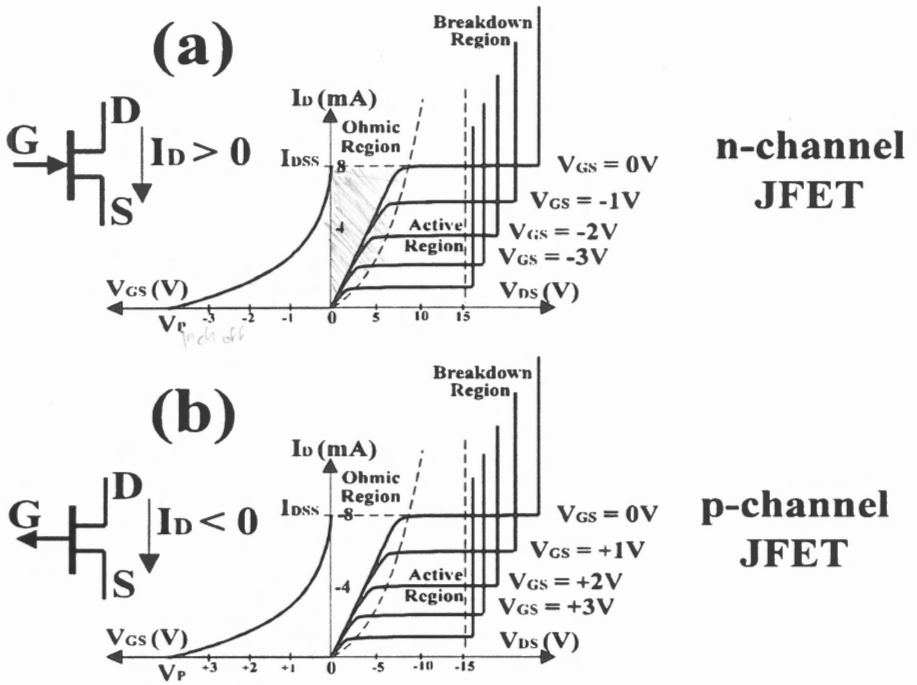


Figure 14.5: n-channel and p-channel JFET Symbols and Characteristics

drain-gate junction and pinch-off occurs. The full set of characteristics, with V_{GS} as the varying parameter, is shown in Figure 14.5. Note that the effects of V_{GS} are reversed in the p-channel JFET.

Figure 14.5 shows the characteristic curves and the circuit symbols for both the n-channel and the p-channel JFETs. Also included in the characteristics graphs is a very useful function called the *transfer characteristic*, which graphs the drain current, I_D, as a function of V_{GS} (with V_{DS} big enough not to matter). The gate-source voltage, V_{GS}, reaches a point at which the channel is pinched off entirely and no current flows at all. This is the *pinch-off voltage*, V_p. At the other extreme, if $V_{GS} = 0$, only the saturation current, I_{DSS} flows (at high V_{DS}). Between these points, the transfer characteristic is roughly parabolic, and is given by the well-known *Shockley equation*:

$$I_D \approx I_{DSS}\left(1 - \frac{V_{GS}}{V_p}\right)^2 \tag{14.3}$$

Hence, knowing only V_p and I_{DSS} (which are given in manufacturers' specifications), the behaviour of a JFET can be quite accurately predicted.

Notice from Figure 14.5 that the JFET may operate in three distinct regions. There is the *ohmic region* in which, as we have seen, it behaves like a resistor of constant value. Secondly, there is the *active region*, in which I_D and V_{DS} are independent, with control of I_D by V_{GS} being possible. This is the region in which the JFET could be used as an amplifier, but note carefully that the dependence of the operating point on V_{GS} is *not linear* because the curves are not equally-spaced for regular variations of the V_{GS} parameter. Furthermore, note that the JFET is a *voltage-controlled* device, unlike the BJT, which may be considered to be controlled by the current, I_B. Finally, note that since pinch-off occurs at varying values of V_{DS}, depending on V_{GS}, the active region of the JFET is somewhat narrower than that of the BJT. The third region of JFET operation, at high values of V_{DS}, arises when the applied drain-source voltage is enough to cause *avalanche breakdown* within the device, leading to a high flow of current. As you can imagine, this is usually undesirable and could easily damage the JFET.

You should see from all of this that JFETs can be used either to switch current on and off (by firmly varying V_{GS}) or for amplification (by applying a small signal to V_{GS} and taking the output as I_D. JFETs can be used in place of BJTs in any of the switching or amplification circuits that we have covered, and the logic circuit, load-line and analysis techniques that we have studied all apply to their use.

14.3 The MOSFET

The JFET works because of the effect of the field arising from the voltage that biases the junction between the channel and parts of the gate. However, there do exist transistors which work entirely on the effects of field, and in which junctions are not important. Figure 14.6 shows both the n-channel and p-channel versions of the *metal oxide silicon field effect transistor* or *MOSFET*, together with the circuit symbols and characteristics graphs for what is known as the *depletion MOSFET*.

In the n-channel depletion MOSFET, current can flow from drain to source along a specially-constructed n-type channel. The aluminium drain and source connectors are laid down on top of *wells*, which form an excellent connection to the channel. Above the channel is a layer of *silicon dioxide* (SiO_2) which is an excellent insulator, and the bottom of the channel is defined by the p-type silicon *substrate*, which is just the silicon base onto which the transistor is built. We can visualise a drain-source current as electrons

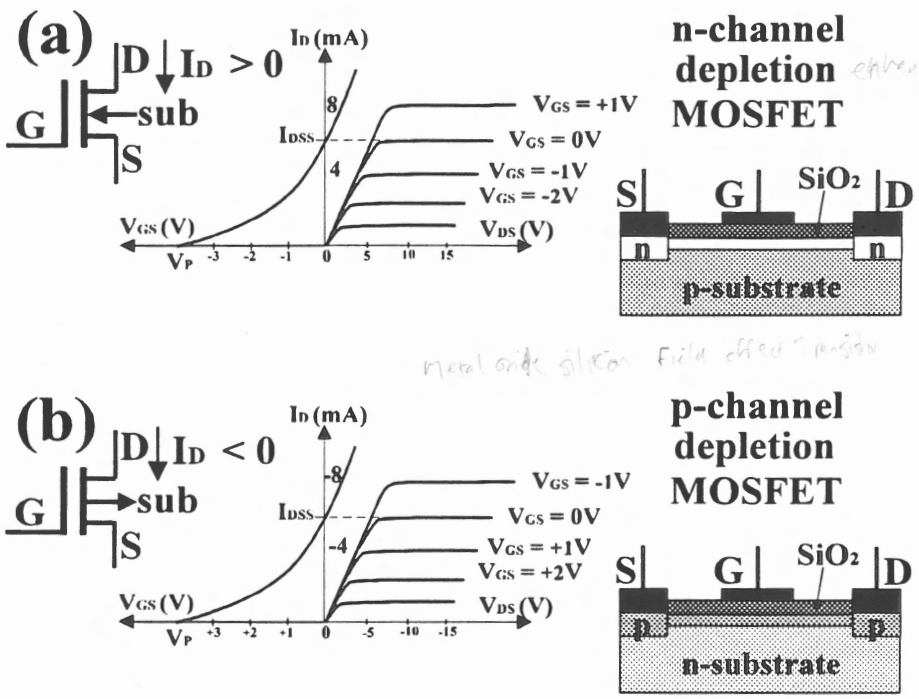


Figure 14.6: The Depletion MOSFET

flowing through the n-type channel from source to drain. With 0V on the gate, it should be clear that, like in the JFET, as V_{DS} is increased there will come a point at which the channel will be saturated and a constant current, I_{DSS}, will flow.

If the gate-source voltage, V_{GS}, is made negative, then the resulting field will push electrons out of the channel, reducing its size by *depleting* it of carriers, and leading to the overall lowering of the I_D-V_{DS} curve that you see in Figure 14.6. You can also make V_{GS} positive in this device, and thus add to the electrons in the channel, increasing its size and raising the characteristic curve. This is called *enhancement*, and MOSFETs which are capable of it are termed *depletion-enhancement* MOSFETs. At some level of V_{GS}, called V_p, the channel is pinched off entirely and no current can flow, while at $V_{GS} = 0$, $I_D = I_{DSS}$ for sufficiently high V_{DS}. The transfer characteristic is therefore very similar to that of the JFET, and it also obeys the Shockley equation. The p-channel depletion MOSFET works in a similar way, but with the polarities of voltages and currents reversed.

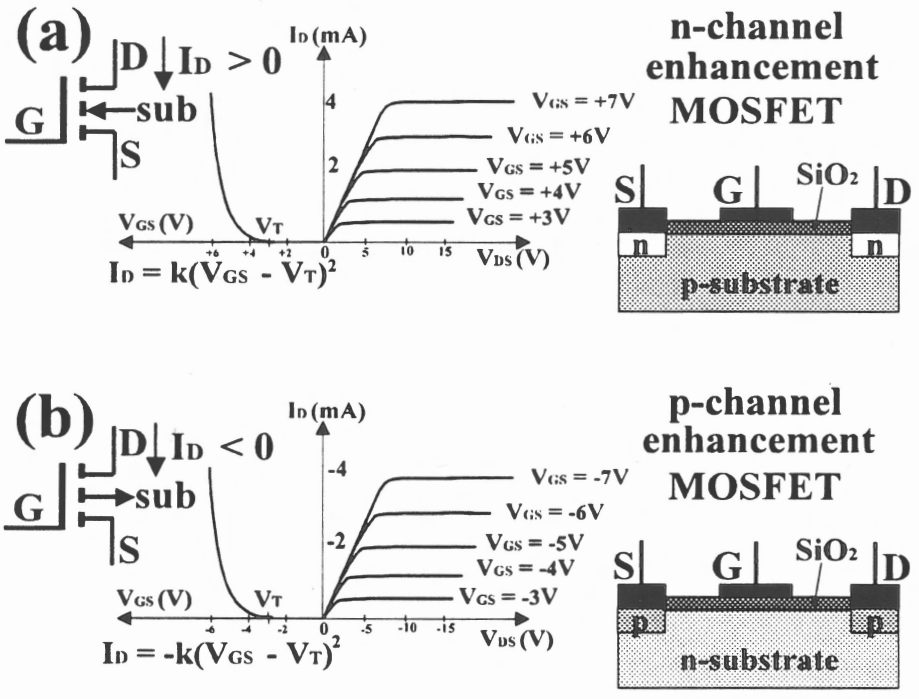


Figure 14.7: The Enhancement MOSFET

Now, the artificial channel in these devices is not strictly necessary to their operation as transistors. In the absence of an n-channel, for example, (but with the SiO_2 insulating layer and n-type wells still present), a positive voltage on the gate would *attract* electrons and thus form a channel of mobile carriers at the top of the p-type substrate. By increasing the positive gate voltage, further enhancement of the channel is possible. Depletion, however, never arises, because there is no channel to deplete! Figure 14.7 shows both the n-type and p-type versions of these *enhancement MOSFETS*, together with their circuit symbols and characteristics. Note the dotted line in the symbols, indicating that the channel is not a permanent feature, but is only created by the gate voltage.

As you see from the characteristics graphs of Figure 14.7, a certain gate voltage is required for any channel to exist at all, and this is called the *threshold voltage*, V_T. Thereafter, drain current rises right up to the avalanche breakdown limit as the gate voltage is increased. There is no saturation effect (i.e. no I_{DSS}) since the channel is not of limited size. Instead, the drain current increases parabolically in response to V_{GS}, with the relationship being given

by

$$I_D = k(V_{GS} - V_T)^2 \tag{14.4}$$

where k is a constant for a particular device. The manufacturer will quote this constant among all the data that are published in the device's printed *specifications*. Typically, k might have a value of 0.3 A/V^2.

Once again, the applications to switching and amplification that these devices offer ought to be apparent to you. If you see them in either rôle, then you should not find it difficult to apply the concepts of load-line analysis that we developed for the BJT, using the characteristics of the FET where appropriate. The main aims of this chapter have been to familiarise you with the various major types of transistor, to give an explanation of how and why they operate as they do, and to show that they are reasonable alternatives to the BJT. In later courses, you will be in a position to appreciate their specific advantages (and disadvantages) more clearly.

Chapter 15

Introduction to Operational Amplifiers

15.1 Integrated Circuits

The last 40 years have seen the building of more and more complicated electronic circuits, and with this has come the need to produce ever-smaller circuit components. As we have seen, diodes and a variety of transistor types can all be constructed out of silicon, provided that the positioning of n-type and p-type regions within the silicon block is carefully controlled. Resistors (up to about 100kΩ) may also be fabricated to quite accurate specifications by making n-type or p-type channels of particular cross-section and length. For a given cross-section, the resistance of the channel is determined by the exact length chosen, and by the doping concentration. There are also techniques for making capacitors in silicon. One method involves using metal or heavily-doped material for the two plates, and then separating them with a thin film of silicon dioxide, which acts as the dielectric. Other methods exploit the fact that pn junctions have capacitance by virtue of the field locked inside them, and that this capacitance can be controlled in size by varying the reverse bias on the junction.

Since all of the principal electronic circuit elements can be made in silicon, the natural development has been to fabricate a whole circuit on a single piece or *chip* of semiconductor. When more than one element is placed on a single chip like this, the resulting device is known as an *integrated circuit* (written *IC*). Over the years, the number of elements seen in integrated circuits has increased dramatically, and you will sometimes see the term *VLSI*

(standing for *very large-scale integration*) that refers to chips that contain over 1000 elements. As you will be aware, the integration of circuits has had a massive impact upon technology, and it continues to revolutionise the way of life of human beings. The miniaturisation and mass-production that it has made possible have brought circuits of ever-greater complexity (and ever-lower price) into the possession of many people.

Although this is a very large subject in its own right, let us take a brief look at *how* integrated circuits are made. From the original circuit diagram, it would have to be decided what configuration of n-type and p-type silicon regions would achieve the same effect. As you can imagine, this decision would require a considerable design effort, particularly for a very large circuit; but with computer-aided design the task is somewhat simplified, and the end-result, once achieved, is highly-repeatable. Even so, the final design will probably call for hundreds or thousands of small layers of n-type or p-type silicon, all to be precisely positioned and sometimes insulated from other parts of the chip by thin layers of silicon dioxide. This extraordinary pattern will be exploiting the conductance properties and the junction phenomena within the silicon chip, to mimic precisely a circuit consisting of many resistors, capacitors, diodes, BJTs and FETs.

The IC is built upon a thick silicon base-layer or *substrate*, which is treated in an evacuated chamber. This chamber can be flooded with gaseous p-type or n-type material, or else with vaporised metal (e.g. aluminium) or insulator (silicon dioxide). Whichever gas is being applied to the chip will tend to be deposited atom-by-atom onto the chip's surface without disturbing the pattern of the lattice structure. This is called *epitaxial vapour deposition.* In this way, it is possible to control the doping types and concentrations, or to lay down layers of insulator or metal contacts, as and where necessary. Often, a layer of silicon dioxide is applied to the whole chip's surface and then minute "windows" are etched in the insulator, using photographic techniques so as to expose certain regions of the underlying silicon before the next layers are added. As you can imagine, there is a vast range of intricate processes associated with IC manufacture, and we cannot hope to cover them here. You would be well-advised, however, to try to read something about IC fabrication, because it will give you a feel for how the principles covered by courses such as this are actually translated into reality.

If you look back at some of the diagrams illustrating BJTs and FETs in earlier chapters, then you should see the intended end-results when these devices are to be put in an IC. Once resistors, capacitors and highly-conductive

channels also became possible on the same chip, it was a small step to the manufacture of the switching and amplification circuits that we have studied. One whole class of ICs is dedicated to realising *logical circuits*, and you will see a great deal of these in your studies in digital electronics. At the moment, we will just mention that, starting from integrated logic *gates* (such as the AND, OR, NOT, NAND and NOR functions), digital circuitry of immense value and complexity has now been developed. This has culminated in the *microprocessors* that are at the heart of all computers and of advanced controlling systems.

15.2 Ideal Operational Amplifiers

For the rest of this course, we will concentrate on integrated *amplifier* circuits, and see some of their many uses in electronics. An early application that they had was to perform mathematical *operations* in analog computers (we shall glimpse how this works shortly). For this reason, integrated amplifying circuits are known as *operational amplifiers* (or *op-amps* for short). A typical op-amp circuit may contain 20 or 30 transistors and many other circuit components. The internal detail of a very well-known op-amp (called the 741) is shown in Figure 15.1. From now on, we shall ignore the components of the op-amp and focus on how the device behaves when it is connected as part of a larger circuit. As you see in Figure 15.1, an op-amp must be supplied from two voltage *rails*, called V^+ and V^-: usually V^+ is a positive voltage and V^- is a negative supply. Apart from the supply voltages, there are two main signal inputs to the op-amp, labelled v_+ and v_-, which are known as the *non-inverting* and the *inverting* inputs respectively. Figure 15.2a shows how both the supply and the signal inputs are usually depicted in op-amp circuits. Note that we now take for granted that all of the elements in Figure 15.1 are included inside the triangular symbol. Figure 15.2b shows a very simple equivalent circuit for an op-amp. The output voltage, v_o, is proportional to the *difference* between the two inputs. This difference is called the *differential input voltage* and is written

$$v_d = v_+ - v_- \tag{15.1}$$

The constant of proportionality is the *gain*, G, so

$$v_o = G(v_+ - v_-) = Gv_d \tag{15.2}$$

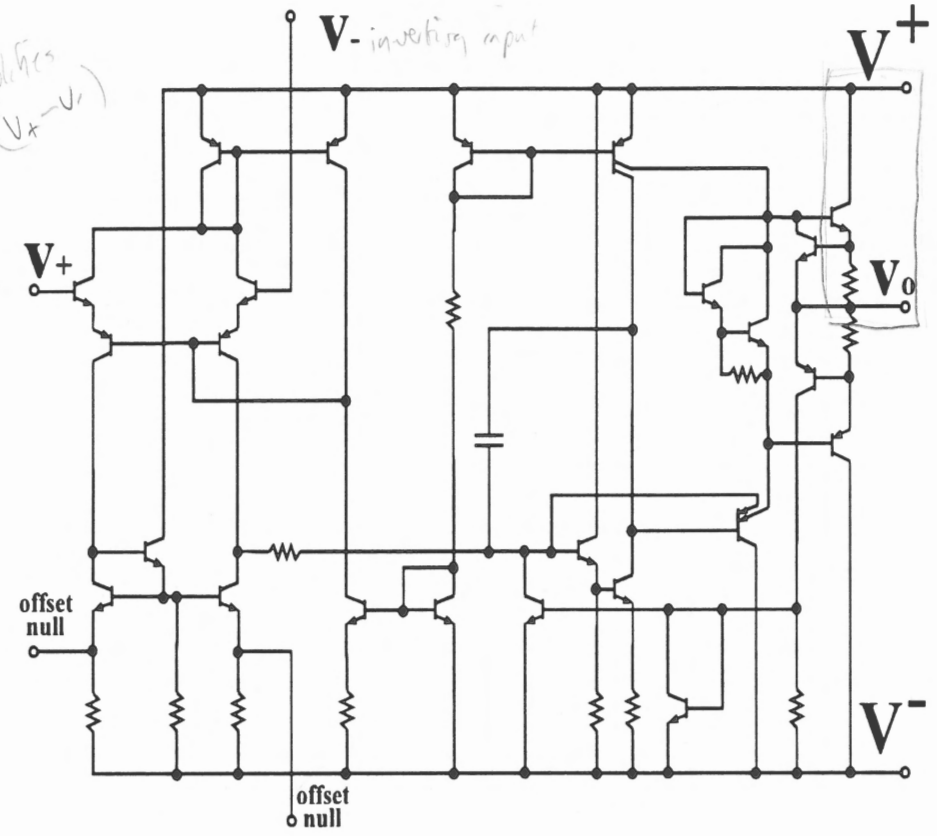
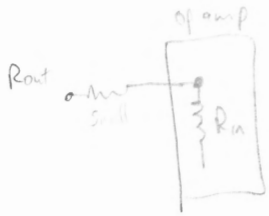


Figure 15.1: The Internal Circuitry of the 741 Op-Amp

Now, if an op-amp were *ideal*, we would have some very stringent expectations of it:



- The input resistance, R_{in}, would be extremely large (in fact, R_{in} would have to tend to infinity) so that the input voltage into the op-amp was not reduced at all by voltage division. (Remember that we assume that whatever supplies the input voltage has an output resistance of its own, which is hopefully small, for the same reason).

- Similarly, the output resistance of the op-amp would be extremely small ($R_{out} \rightarrow 0$) to ensure that the output voltage was not divided down as it entered the next stage of the circuit, following the op-amp. (Remember that the next stage will have its own input resistance, which is hopefully large).

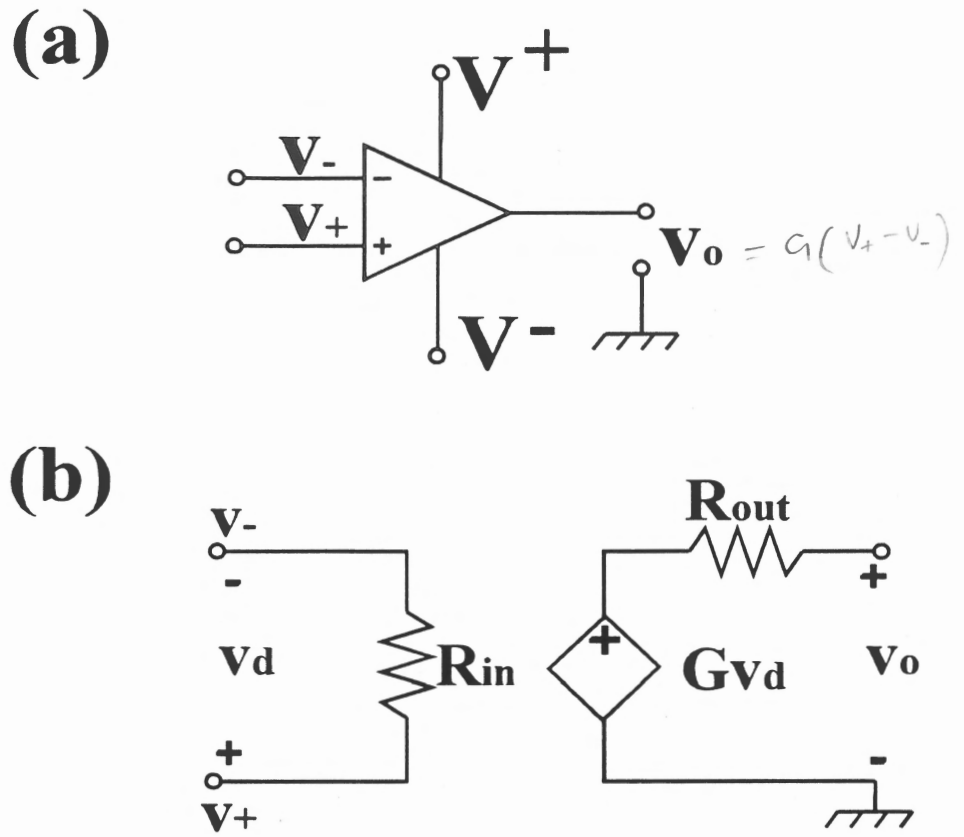


Figure 15.2: The Op-Amp Symbol and Equivalent Circuit

- The gain of the op-amp would be huge (i.e. it would amplify perfectly). Ideally, $G \rightarrow \infty$. In practice, $G \approx 10^5$.
- The op-amp would amplify the difference voltage, no matter how fast it may vary. In the jargon, it should have *infinite bandwidth.*
- In spite of the infinite gain, if $v_+ = v_-$, then the output of the ideal op-amp should be $v_o = 0$.

If you think about these conditions, then you will see that they are each impossible separately and, indeed, that they would also be impossible taken together, even if achievable separately! Nevertheless, they represent an *ideal* device which it is important to understand well.

Notice how the gain equation gives us

$$v_+ - v_- = \frac{v_o}{G} \tag{15.3}$$

so that, as $G \to \infty$, we get

$$v_+ - v_- = 0 \qquad so \qquad v_+ = v_- \tag{15.4}$$

This gives us our first op-amp Rule:

Ideal op-amps tend to drive their two input voltages to equality

Then, since the input resistance is infinite, Figure 15.2b makes it clear that the current into the two terminals, by Ohm's Law, is zero. Hence, our second Rule:

Ideal op-amps draw no current at either input terminal

To these let us add, as the third Rule, a statement about amplification:

The output of the ideal op-amp is equal to the differential input voltage times the gain (or $v_o = Gv_d$)

These three simple Rules give you all that you need to analyse circuits containing ideal op-amps. As we shall see, they also perform very well for real circuits containing practical op-amps. This chapter has aimed to give a brief introduction to what an op-amp is: your aim after reading it should be to ensure that you are comfortable with these three Rules. They will be our fundamental analysis tool in the coming chapters.

Chapter 16

The Inverting Amplifier

16.1 Feedback and Closed-Loop Gain

Real op-amps, connected as shown in Figure 16.1a, have their main use as *comparators*. If the non-inverting input terminal voltage, v_+, is greater than the inverting terminal voltage, v_- (even by a small amount), then the output voltage will be large and positive, since $v_o = Gv_d$ and $G \approx 10^5$ (see Figure 16.1b). Conversely, if v_+ is less than v_-, then the output voltage will be large but negative. In practice, since the op-amp is supplied from *rails* at voltages of V^+ and V^-, the output cannot exceed these limits, and will *saturate* somewhere near the rail values. So the comparator will have $v_o \approx V^+$ if $v_+ > v_-$ and $v_o \approx V^-$ if $v_+ < v_-$. You will encounter this very useful device often in your later work.

For the moment, notice that since the circuit of Figure 16.1a has v_- connected to ground ($v_- = 0$V, since almost no current flows through the resistor R into the v_- terminal), this circuit can *test* any voltage placed on v_+ to see if it is positive or negative. Its output will be a large voltage of the same polarity as the input v_+. If we had grounded v_+ instead of v_-, then the output would have the opposite polarity to any input placed on v_-, and we should say that the comparator was *inverting*. In this case, the circuit would have negative gain, but we could take that into account and still use it as a perfectly reasonable test for the input's polarity.

We call the configuration of Figure 16.1 an *open-loop connection*, and the gain, G, is termed the *open-loop gain*. It is much more common, however, to see an op-amp connected as in Figure 16.2, which shows a *closed-loop* connection. In this circuit, there is a path for some current to flow from the

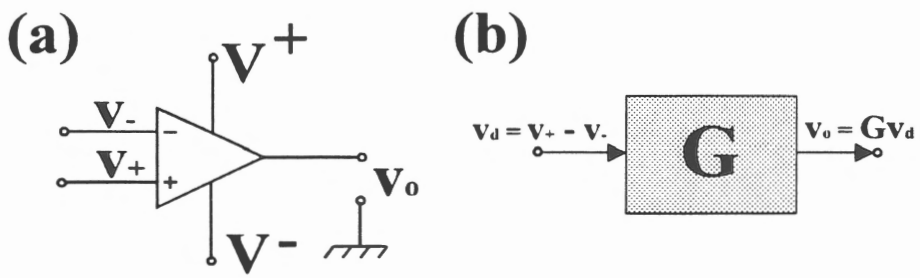


Figure 16.1: Open-Loop Connection: Useful as a Comparator

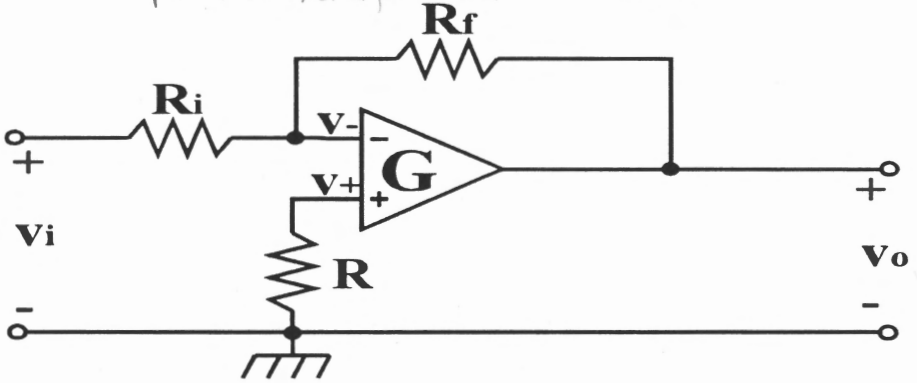


Figure 16.2: Closed-Loop Connection: Negative Feedback

output *back* into the inverting input terminal of the op-amp - a situation which we call *negative feedback*. Now, feedback is a very valuable mechanism in engineering, and it gives rise to many subtle ideas. However, as it is easy to be put off by its more difficult aspects, we will concentrate at present on demonstrating three essential points about the circuit of Figure 16.2:

- It is an *inverting amplifier*
- Its *closed-loop gain* is ideally $-R_f/R_i$
- It will perform almost exactly the same in practice, with a *real* op-amp connected, as it does in theory, when the op-amp is assumed ideal.

That the circuit of Figure 16.2 is *inverting* is strongly suggested by the fact that v_+ is grounded (through a resistor) while signals enter the op-amp through v_-. Since (by our second op-amp Rule) no current flows into v_+,

evidently $v_+ = 0$ and so

$$v_o = Gv_d = G(v_+ - v_-) = -Gv_- \tag{16.1}$$

Thus the circuit has an *open-loop* gain of G, and the negative sign denotes that it is inverting.

To find the overall gain (the *closed-loop gain*) between the input v_i and the output v_o of the circuit, let us write the nodal equation for the v_- terminal:

$$\frac{v_- - v_i}{R_i} + \frac{v_- - v_o}{R_f} = 0 \tag{16.2}$$

Notice, again, the assumption that no current flows into the v_- terminal. Our first op-amp Rule now allows us to set

$$v_+ = v_- = 0 \tag{16.3}$$

When v_- is driven to 0V by op-amp action, it is termed a *virtual earth*, meaning that it is at the same potential as the ground, but that no current flows. The last two equations now quickly give us

$$\frac{v_o}{v_i} = -\frac{R_f}{R_i} \tag{16.4}$$

This is a very significant result, because it means that we can convert an op-amp with a gain of G into an (inverting) amplifier with *any* gain that we please, determined by the values of the connected resistors, R_i and R_f.

In deriving this result, however, we used all of the three Rules which hold for *ideal* op-amps. What can we say about the amplifier circuit when we try to build it using a real op-amp? As long as the input resistance of the op-amp, R_{in}, is very large (2MΩ is a typical value for the 741 op-amp) and perhaps if the resistor down to ground, R, could be made large as well, Ohm's Law tells us that very little current will flow into the non-inverting terminal, and so we can write

$$v_+ \approx 0 \qquad \textit{and} \qquad v_o = Gv_d \approx -Gv_- \tag{16.5}$$

If you now take the result, $v_- \approx -v_o/G$, and substitute it into the nodal equation for the v_- terminal, you will get

$$\frac{-v_o/G - v_i}{R_i} + \frac{-v_o/G - v_o}{R_f} \approx 0 \tag{16.6}$$

which, after some algebra (which you ought to check) becomes

$$\frac{v_o}{v_i} \approx -\frac{R_f}{R_i}\left[\frac{G}{R_f/R_i + 1 + G}\right] \tag{16.7}$$

In a practical op-amp such as the 741, a typical value for G is about 200000, so as long as R_f and R_i are within a factor of (say) 1000 of each other, we can agree that G dominates the fraction in the above expression, and we can write

$$\frac{v_o}{v_i} \approx -\frac{R_f}{R_i} \tag{16.8}$$

This, of course, is what we saw predicted in the ideal case, and it confirms that circuits built with real op-amps will perform as the ideal theory suggests. Note how this means that the closed-loop gain of the inverting amplifier is *independent of* G. Having satisfied ourselves as to the relevance of the theory, we can now use it to analyse some interesting op-amp-embedded inverting amplifier circuits.

16.2 Analysis of Inverting Amplifier Circuits

The process of analysing op-amp circuits with feedback is refreshingly easy! Many slightly different configurations exist, which perform a variety of functions, as we shall see; but in each case, the analysis follows a similar pattern and you can rely upon your ideal op-amp Rules to predict the behaviour of actual circuits.

Let us start by looking at Figure 16.3, which shows an inverting amplifier, amplifying a signal, v_i, and supplying a load, R_L. We wish to find the output voltage across the load. One method is to spot that $v_- = v_+ = 0$ (by Rules 1 and 2) and then to write the nodal equation for v_-:

$$\frac{v_- - v_i}{R_i} + \frac{v_- - v_o}{R_f} = 0 \tag{16.9}$$

from which

$$v_o = -\frac{R_f}{R_i} v_i \tag{16.10}$$

Alternatively, after identifying the virtual earth, $v_- = 0$, you can find the current through the input resistance,

$$i_i = \frac{v_i - 0}{R_i} = \frac{v_i}{R_i} \tag{16.11}$$

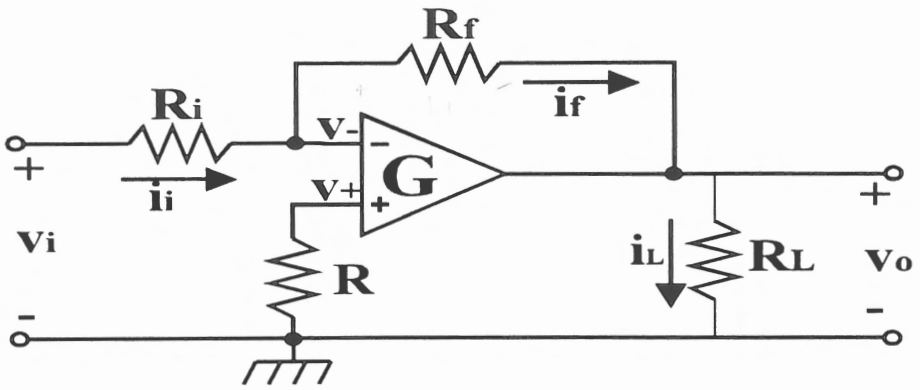


Figure 16.3: An Op-Amp-Embedded Inverting Amplifier

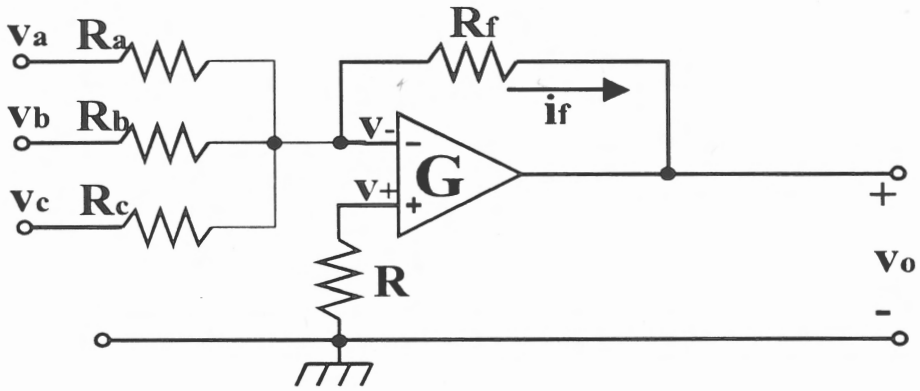


Figure 16.4: An Inverting, Weighted Summer

and then use Rule 2 to argue that since (almost) none of this current passes into the v_- terminal, it must all flow through R_f. Thus $i_f = i_i$. The voltage *drop* across R_f is therefore

$$v_f = i_f R_f = \frac{v_i R_f}{R_i} \tag{16.12}$$

Hence

$$v_o = 0 - v_f = -\frac{v_i R_f}{R_i} \tag{16.13}$$

which is the same answer as before. The circuit of Figure 16.3 is a simple inverting amplifier with a gain of R_f/R_i.

Now look at Figure 16.4, into which inputs v_a, v_b and v_c are connected.

Clearly, $v_- = v_+ = 0$ and nodal analysis gives us

$$\frac{v_- - v_a}{R_a} + \frac{v_- - v_b}{R_b} + \frac{v_- - v_c}{R_c} + \frac{v_- - v_o}{R_f} = 0 \qquad (16.14)$$

Combining these facts and re-organising the nodal equation gives us

$$v_o = -\frac{R_f}{R_a}v_a - \frac{R_f}{R_b}v_b - \frac{R_f}{R_c}v_c \qquad (16.15)$$

Now, if $R_a = R_b = R_c = R_f$, then we simply have

$$v_o = -(v_a + v_b + v_c) \qquad (16.16)$$

In other words, the circuit is an *inverting summer*, whose output voltage is just the (inverted) sum of the three input voltages. If, however, we had $R_a = R_b = R_c$ but R_f were some greater resistance, then we should have the sum of the inputs, inverted and amplified, appearing at the output. In the case where $R_a \neq R_b \neq R_c$, each input is *weighted* by a different amount (R_f/R_a, R_f/R_b and R_f/R_c respectively), and then they are inverted and summed at the output.

Other interesting variants of these circuits are possible, and you should try experimenting to see what oher effects can be achieved. You also ought to make sure that you could draw the circuit diagram for any inverting summer or amplifier. if a desired output is specified. For example, could you draw a circuit with inputs v_a and v_b whose output was $v_o = -(v_a + 10v_b)$?

Let us finish with the circuit of Figure 16.5, into which v_1 and v_2 are inputs. The voltage at the non-inverting terminal, v_+, is $v_2/2$ by simple voltage division, so v_- must also be $v_2/2$, by op-amp action (Rule 1). The current flowing from the v_1 input is simply $(v_1 - v_2/2)/R'$, by Ohm's Law. Thus. again by Ohm's Law,

$$v_o = \frac{v_2}{2} - \frac{v_1 - v_2/2}{R'}.R' = v_2 - v_1 \qquad (16.17)$$

The circuit of Figure 16.5 is therefore an unweighted *subtractor*. This should open up another range of possibilities for you to explore!

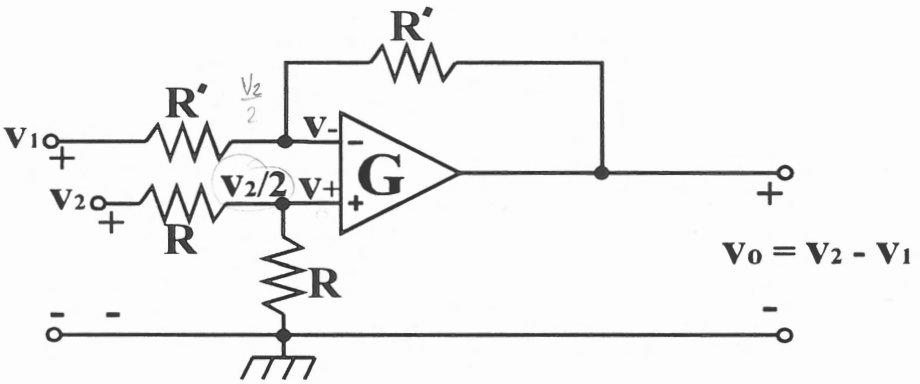


Figure 16.5: An Unweighted Subtractor

Chapter 17

Non-Inverting Amplifiers

17.1 Calculating the Gain

The subtractor with which we finished the last chapter is not strictly an inverting amplifier, because it has an input at the non-inverting terminal as well as at the inverting terminal of the op-amp which it contains. In this sense, it combines the features of inverting and non-inverting circuits. We now concentrate on circuits like the one in Figure 17.1, in which the input, v_i, is into the non-inverting terminal only. We shall soon see that the output of such circuits has the same polarity as the input, and that there is generally a gain factor. For this reason, we call the circuit a *non-inverting amplifier*. Note that Figure 17.1a and Figure 17.1b are both drawings of the *same* circuit: they are both included here to help you to recognise a non-inverting amplifier circuit, whichever way it may be drawn.

Analysing the non-inverting amplifier circuit follows familiar and easy lines. First, you write an equation for the v_+ node, using our op-amp Rule 2 to get

$$v_+ = v_i \tag{17.1}$$

and then you write the nodal equation for the v_- terminal:

$$\frac{v_-}{R_a} + \frac{v_- - v_o}{R_f} = 0 \tag{17.2}$$

Rule 1 now gives you $v_+ = v_- = v_i$, so we can replace v_- with v_i in the above equation and then re-organise it, to obtain the (non-inverted) gain of the circuit:

$$\frac{v_o}{v_i} = 1 + \frac{R_f}{R_a} \tag{17.3}$$

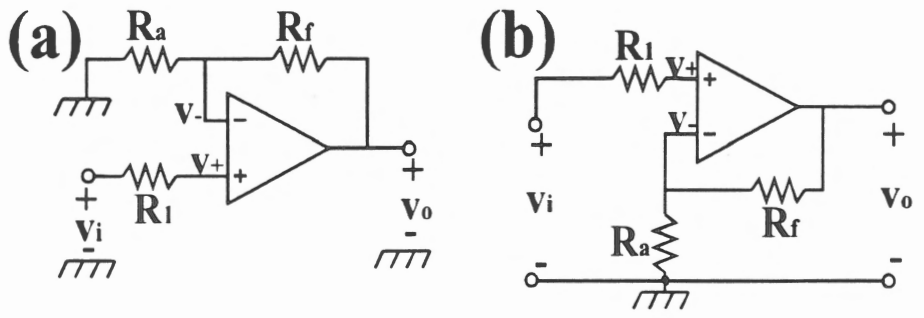


Figure 17.1: Two Views of a Non-Inverting Amplifier

Now, you ought to be aware that the input to an op-amp-embedded amplifier does not need to be a dc voltage. On the contrary, one often wishes to amplify a signal that *changes*, perhaps sinusoidally, but perhaps irregularly (like the output from a microphone, if you talk into it).

Let us have a look at the slightly more complicated amplifier circuit of Figure 17.2. The idea here is for the sinusoidal input signal to be amplified and then applied to the 2kΩ load. We will calculate the current delivered to the load by the amplifier. Since there is a resistor to ground from the non-inverting terminal, we begin by finding the voltage there, v_+, using simple voltage division. Thus,

$$v_+ = \frac{20}{20+10} v_i = 333 \sin(\omega t)\ \text{mV} \tag{17.4}$$

and, since $v_- = v_+$, we can apply voltage division again (in reverse) to get

$$v_o = \frac{98+7}{7} v_- = 5 \sin(\omega t)\ \text{V} \tag{17.5}$$

This result could, of course, have been reached using the method by which we analysed Figure 17.1. It is now an easy step to use Ohm's Law to get

$$i_L = \frac{v_o}{R_L} = 2.5 \sin(\omega t)\ \text{mA} \tag{17.6}$$

Both this current and the small current that flows in the feedback loop and then to ground must come from the output terminal of the op-amp. If you look back at Figure 15.1, then you can see how the current must come from the power rail of the op-amp. From there, it must flow through the transistors and resistors of the *output stage* before leaving the op-amp to go either to the load or to the feedback loop. We do not want to get too

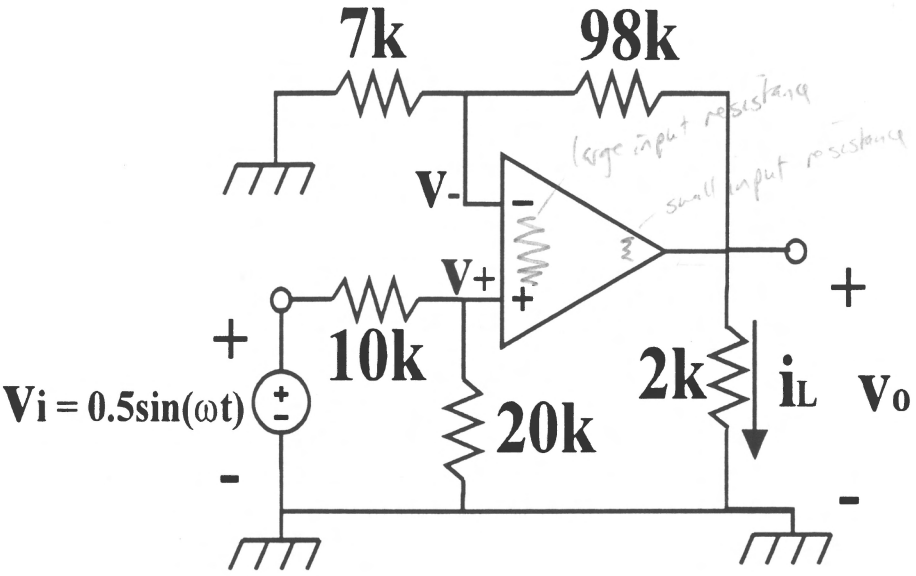


Figure 17.2: A Non-Inverting Amplifier Drives a 2kΩ Load

involved with the internal workings of the op-amp, but it ought to be clear that, since the rail voltage (V^+) is finite and since there are small resistors in the output stage, the current that the op-amp can supply is limited. In fact, even if we short-circuited the op-amp output in Figure 17.2, a current of no more than about 500mA would flow. However, the manufacturers of the op-amp will specify a *maximum short-circuit output current* (it is about 40mA for the 741), and if this is exceeded then the op-amp may be damaged. Hence, one should never short-circuit an op-amp output, or use a load with such a low resistance that the output current exceeds about 40mA.

17.2 Input and Output Resistance

You may wonder why the manufacturers of the op-amp don't just put bigger resistors into the output stage, so that there is no danger of drawing a current that will damage the op-amp. If they did that, then there would be a loss of gain because the op-amp output resistance and the load resistance would present a voltage divider to the op-amp output voltage, thereby diminishing the voltage across the load. You may recall from our discussion of ideal op-amps that this is why manufacturers strive to make the *output resistance as small as possible.*

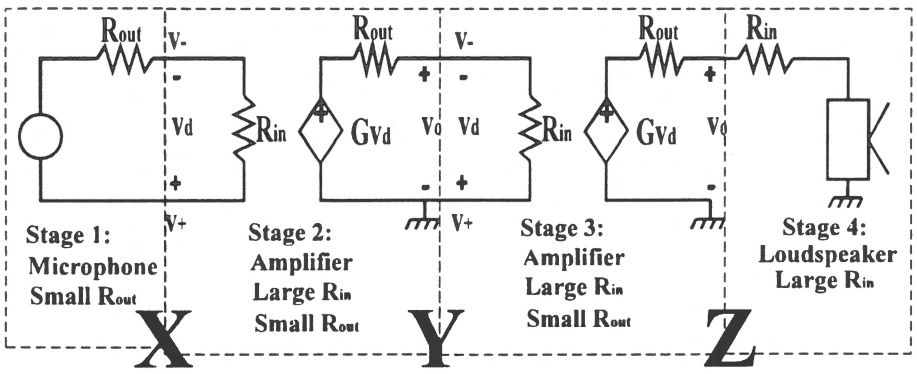


Figure 17.3: An Ideal Multi-Stage Amplifier Circuit

As for the op-amp's *input resistance*, this has to be *as large as possible*, so that the op-amp will not draw excessive current from whatever is feeding its inputs and hence cause damage or undesirable voltage division (a problem called *loading*). The input applied to the op-amp, whether it be from a microphone, from a voltage supply or even from the output stage of another op-amp, has got an output resistance of its own, and this, hopefully, will be *small* (again to minimise the effects of voltage division). A *multistage* circuit, summarising the ideal situation, is shown in Figure 17.3, in which the op-amp stages have been replaced by their equivalent circuits so that you can clearly identify their input and output resistances. Note that the whole idea of having two op-amp stages would be to achieve greater amplification of the input signal (from a microphone in this case) before delivering the signal to the output (a loudspeaker).

Now, a real op-amp has a very high input resistance but, regrettably, an inverting op-amp-embedded amplifier circuit does not. With the feedback loop now added, the amplifier has an equivalent circuit shown in Figure 17.4. Suppose that we input a voltage v into this circuit and a current i then flows into it as shown. KVL gives us

$$v - i(R_a + R_f) - G(v_+ - v_-) = 0 \tag{17.7}$$

But since

$$G(v_+ - v_-) = Gv_d = -G(v - iR_a) \tag{17.8}$$

we get

$$i(R_a + R_f) = (1 + G)v - GiR_a \tag{17.9}$$

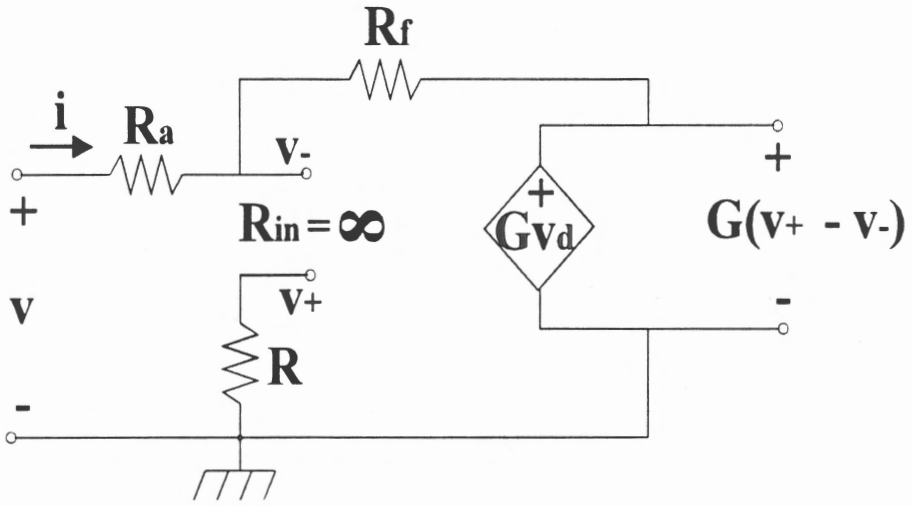


Figure 17.4: Equivalent Circuit For an Ideal Inverting Amplifier

Ohm's Law now tells us that the input resistance of the amplifier is just

$$R_{input} = \frac{v}{i} = \frac{R_a + R_f}{1+G} + \frac{GR_a}{1+G} \tag{17.10}$$

We know that the open-loop gain, G, is of the order of 10^5 for a real op-amp, so provided that R_a and R_f are no bigger than a few kilohms, we can afford to ignore the first term in this expression. The second term, however, will not go away, and we must conclude that, for the *inverting* amplifier

$$R_{input} \rightarrow R_a \tag{17.11}$$

17.3 Buffers

This all suggests that the circuit of Figure 17.3 is impossible in practice, because the op-amp stages do *not* have the high input resistance that we now know is vital. We need to look for a way to make this circuit work so that its various stages do not *load* preceding stages. Consider Figure 17.5, which shows a non-inverting op-amp configuration with a single input to the v_+ terminal. The only connection to the v_- terminal is to the feedback loop from the output. Our op-amp Rules very quickly show that

$$v_i = v_+ = v_- = v_o \tag{17.12}$$

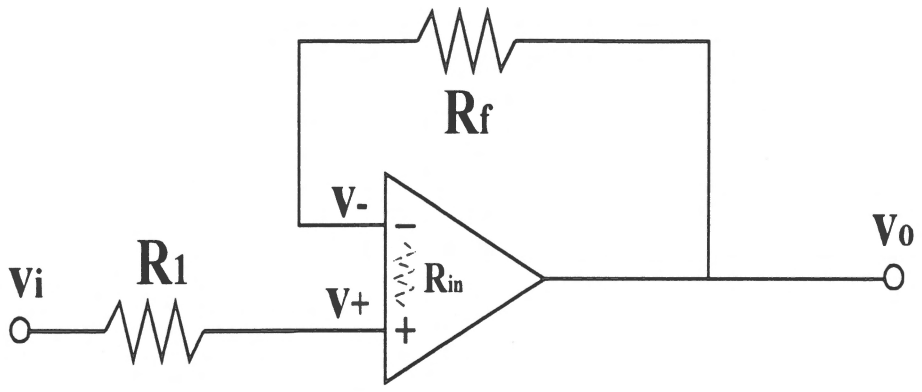


Figure 17.5: A Unity Gain Buffer

So we have *unity gain*, and the circuit apparently gets us nowhere! But look again!! The input resistance to this circuit is just a series of resistors, including R_{in}, which is very *large*:

$$R_{input} = R_1 + R_{in} + R_f \tag{17.13}$$

while the output resistance has only the *small* value that we associate with the op-amp output stage (further reduced by feedback). The circuit of Figure 17.5 is known as a *unity-gain buffer*. Evidently, if we simply inserted it into the points marked X, Y and Z in Figure 17.3, we would obtain a multistage amplifier in which loading was no longer a problem. Note also that by varying the resistances R_f and R_1 in the buffers, we can get some further gain out of them without compromising their buffering function.

We complete this chapter by looking at the circuit of Figure 17.6, which shows a *non-inverting summer with gain.* It gives us an example of how operations such as voltage addition can be performed by non-inverting op-amp-embedded circuits. At the v_+ terminal we have

$$\frac{v_+ - v_1}{R_1} + \frac{v_+ - v_2}{R_2} = 0 \tag{17.14}$$

while at the v_- terminal, voltage division simply gives

$$v_- = \frac{R_a}{R_a + R_f} v_o \tag{17.15}$$

If we set $v_+ = v_-$, then we obtain after some algebra:

$$v_o = (R_1 // R_2)\left(\frac{v_1}{R_1} + \frac{v_2}{R_2}\right)\left(1 + \frac{R_f}{R_a}\right) \tag{17.16}$$

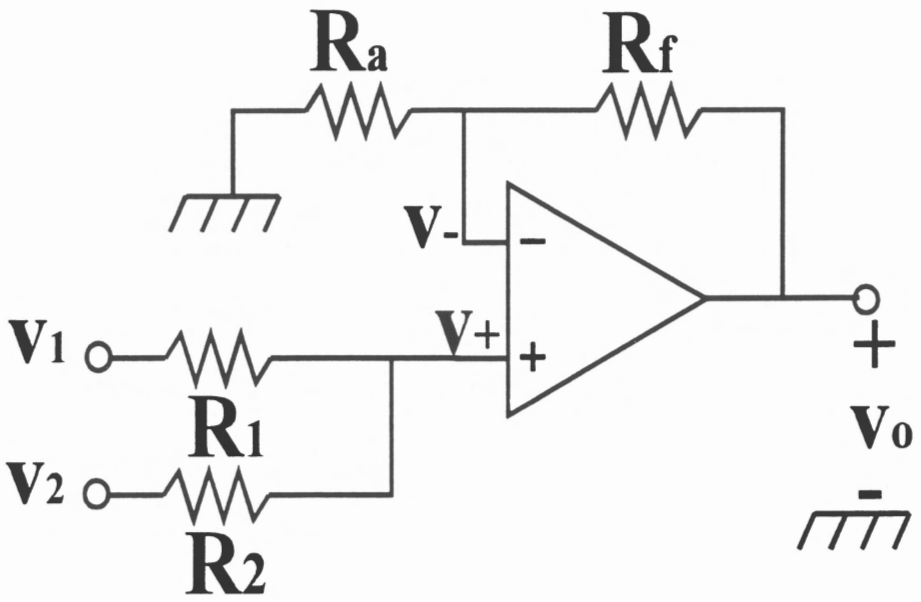
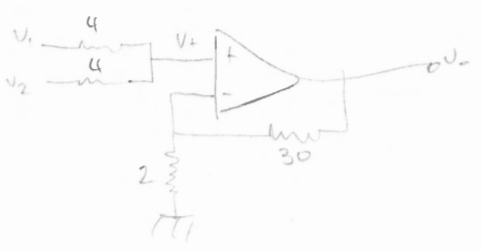


Figure 17.6: A Non-Inverting Summer With Gain

You should confirm this result and check that you agree that, in the special case where $R_1 = R_2$ and $R_f = R_a$, we would obtain $v_o = v_1 + v_2$.

Notice that, unfortunately, multiple-input non-inverting amplifier circuits do not possess buffering abilities. In Figure 17.6, the input resistance "seen" by an ac signal at v_1 or v_2 is $R_1 + R_2$. If there were *three* inputs, then the resistance "seen" at v_1 would be $R_1 + R_2//R_3$. You can easily extend this to further cases with more inputs.

E 3.5.6



$$\frac{R_a + R_f}{R_1 + R_2}\left[\frac{R_1}{R_a}V_2 + \frac{R_2}{R_a}V_1\right] = \alpha V_o$$

Chapter 18

Sensors & Transducers (1)

18.1 Some Measurement Concepts

We have talked a lot about circuits that *process* signals by combining them in logical switching operations or by amplifying them. We have not, until now, concerned ourselves with where those signals originated. A major branch of electrical engineering is the field of *electronic measurement*, where the main aim is to detect and quantify a change in a physical variable in the world around us, and then to convert that change to an electronic form, which is capable of being switched, amplified, displayed or stored. Many devices have been developed that are capable of detecting a whole range of such physical variations, and this chapter and the next will look briefly at some of them. The variety of possible applications is vast, and one can obtain devices that *sense* light, temperature, sound, magnetic fields, mechanical forces, position, proximity, angle, humidity, acidity, gases and nuclear radiation. Of course, this introduces a field that you could study for years all by itself, so the aim here is just to get the flavour of it by describing a few sensing devices and mentioning the physical principles upon which they operate.

Just before we begin, we shall need a few basic definitions. A *sensor* is a device that responds to a change in a measured physical quantity (called the *measurand*) by producing a change in a physical property of its own. We are interested in *electrical* sensors, so the *output* of the sensor will usually be an electrical property, such as a voltage or current, or possibly the resistance or capacitance of part of the sensor. You often see the word *transducer* used interchangeably with *sensor*. Strictly, a transducer is a device which converts one form of *energy* into another form of energy. Thus, an electrical sensor

which detects light and outputs a voltage (without any need for a power supply) is converting light energy to an electrical form, and is therefore also a *transducer*. However, a device whose resistance depended on the level of light falling upon it could be a useful *sensor* of the light level, but is *not* a transducer, because no energy conversion takes place within it.

A sensor (or transducer) will usually have certain basic properties:

- It will hopefully be strongly affected by the physical quantity that it is designed to measure, but will not react to other physical quantities. For example, we expect a microphone to be *sensitive* to sound, but its performance should *not* depend on its temperature, humidity etc.
- The amount of change in the sensor's output should be related to the amount by which the measurand changes (e.g. the higher the measurand, the higher the sensor's output). It is particularly useful if they can be related in *fixed proportion*, in which case we call the device a *linear sensor*
- There will be a certain small amount of change in the measurand which is just too small to cause a change in the output. This represents the *sensitivity* limit of the device.
- There will be extremes of the measurand beyond which the device will be damaged or destroyed (think of an electronic thermometer, for example) but within which it is often assumed to exhibit good, sensitive, linear behaviour

Armed with these simple generalisations, we are ready to explore some of the many interesting devices that are available.

18.2 Radiation Sensors

Radiation sensors respond to changes in electromagnetic radiation in their vicinity, and you can get sensors which are sensitive to various specific portions of the electromagnetic spectrum. At the longer wavelengths, radio waves are detected using circuits which *resonate* at a desired frequency - this, of course, is the principle of the *radio receiver*. With decreasing wavelength, other sensors exist which respond to infra-red (IR), visible light or even ultra-violet (UV) radiation.

Perhaps the simplest radiation sensor is the *photovoltaic cell*, sketched in

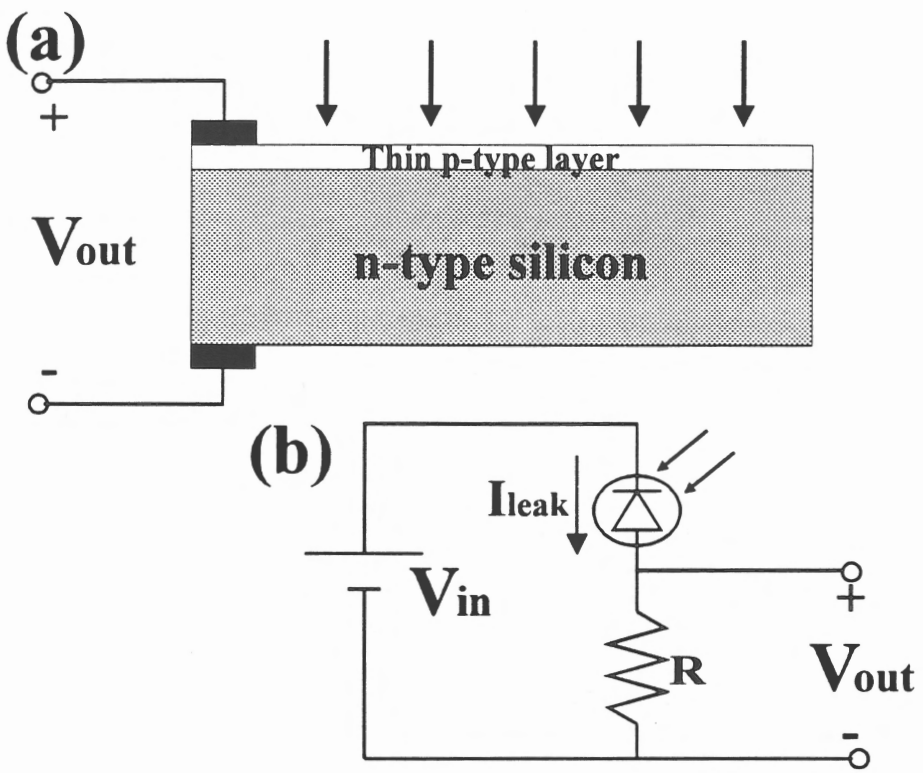


Figure 18.1: (a) A Photovoltaic Cell (b) A Photodiode

Figure 18.1a. This is simply a pn junction of very large area (a typical device is 100mm^2) and with a very thin upper layer (the p-type layer in this case). As with any silicon pn junction, a depletion layer automatically forms and a 0.7V *virtual cell* is created. When radiation falls onto the device, some electrons escape from the atoms of the semiconductor and leave mobile holes in the crystal structure. The embedded field in the depletion layer attracts the holes to the p-type side and the conduction electrons to the n-type side, from which they may then flow into any external circuitry. A typical *responsivity* of such a device would be 0.5 amps per watt of radiant energy received, varying linearly with the incident radiation. As long as radiation falls on the cell, the supply of electrons and holes is continually renewed, generating an output voltage as shown. No external power supply is needed, so the device is a true transducer. Its peak sensitivity is at 800-900nm (in the red/IR part of the spectrum), but colour filters can be incorporated into the front of the device as a window, allowing the peak sensitivity to be set as desired within

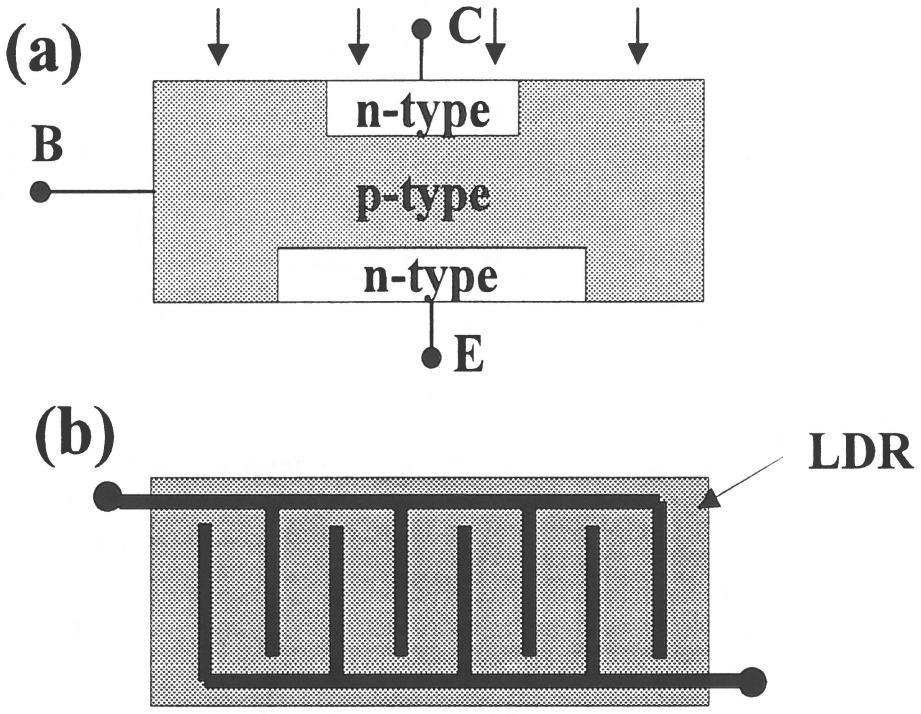


Figure 18.2: (a) A Phototransistor (b) A Light-Dependent Resistor

the visible spectrum. Note that *solar cells* have very similar construction and performance: you will have seen these used as power-generating devices in remote areas and on spacecraft.

Figure 18.1b shows a *photodiode*, which is a light sensor that we met earlier in the course. You may recall that these are connected to a power supply under reverse bias. Many photodiodes are simply ordinary diodes encased in clear glass intead of in an opaque package. When light enters the device, although majority carriers cannot cross the potential hill at the pn junction, *minority carriers* are created (electrons in the p-type and holes in the n-type) which are swept across the depletion layer to form an output current. With a resistor connected as shown, the output may be taken as the voltage across the resistor. Again, the response of these devices is linear, but because their output is due to the movement of minority carriers, the size of the output is rather limited.

This problem is overcome in the *phototransistor*, a device illustrated in Fig-

ure 18.2a. This is a BJT with a transparent lens that allows light to reach the base layer. If the BJT is an npn construction, then the collector-base junction is reverse-biased. Light on the base produces electron-hole pairs at this junction, allowing a minority carrier current to flow. Transistor action then causes a much greater current to pass from the collector to the emitter. You can therefore regard the phototransistor as simply a photodiode (the collector-base diode) with built-in amplification. The output might be taken as the collector voltage or as the voltage across some externally-connected resistor. Although the phototransistor is more sensitive to light than the photodiode, it takes longer to respond to changes in light intensity (milliseconds instead of nanoseconds). This might be a factor against using it in certain high-speed applications.

The final radiation sensor that we shall consider is the *light-dependent resistor* (or *LDR*) shown in Figure 18.2b. As we know, it is a feature of all semiconductors that if energy is added (at an appropriate frequency) then carriers are excited into the conduction band, and electron-hole pairs are created. Certain semiconductors, such as cadmium sulphide or cadmium selenide, are known to produce minority carriers particularly copiously if light falls upon them. Lead sulphide and mercury cadmium telluride also have this property. With incident light, so many carriers are liberated in these substances that their ability to conduct electrical current is greatly increased (i.e. their resistance is greatly reduced). An LDR consists of a block of such material separating a pair of conductors. To maximise the sensitivity to light, the conductors are made to be *interdigitating*, as shown in Figure 18.2b. The result is a resistor which varies from about 10MΩ in darkness to around 100Ω in bright light. These devices are *not* linear, and they are very slow, taking up to 1/3 second to respond fully to a light intensity change. However, they are very cheap, easy to use and highly sensitive. In a well-designed circuit, they will give an excellent response to the presence or absence of surrounding light.

18.3 Thermal Sensors

The measurement of temperature brings with it its own special problems, and a range of devices exist which overcome these. As you will know, heat may be transferred by *conduction* or *convection*, or else by *radiation* of electromagnetic energy from the hot source. If you want to measure the temperature of something very hot or of something which you cannot approach

(such as the inside of a blast furnace, the surface of a distant star or the exhaust gases of a departing rocket) then you have to rely on radiated energy to give you the information that you want. Devices which measure temperature by responding to radiated energy are called *pyrometers*, and many types exist, including some that are based on the photovoltaic devices of the previous section. In brief, pyrometers usually work by pointing a small telescope at the region of interest, and then separating the infra-red radiation seen through the telescope from the other frequencies. The intensity of the infra-red being received is a measure of the temperature of the object being viewed, and this would be taken as the output of a photovoltaic sensor, placed at the end of the telescope.

More conventionally, one measures temperature by measuring the heat emitted from an object by conduction or convection, and for this there are also various devices available. A *thermistor* is a thermally-sensitive resistor (usually made from oxides, sulphides or selenides of nickel, manganese, cobalt, copper or iron) whose resistance decreases as the surrounding temperature increases - that is, the material of the resistor has a *negative temperature coefficient*. At a reference temperature, such as $T_{ref} = 300\text{K}$, the resistance R_{ref} of the resistor is well-known: thereafter, at any temperature, T in Kelvins, its resistance can be found as

$$R_T = R_{ref} e^{-\beta(1/T - 1/T_{ref})} \tag{18.1}$$

where β is a parameter that depends on the material of the resistor. As you can see, thermistors are *not* linear sensors, but they do have some compensating advantages. Firstly, they work well over a temperature range from -50°C to +300°C, and secondly they can be made very small. This means that they can respond quickly to changing temperatures, and that they are particularly convenient to attach to or insert into objects, animals or people whose temperature is to be monitored. A major disadvantage of thermistors is that they require a power supply, and the passage of current through the device causes a little *self-heating*, which makes very precise temperature measurements impracticable.

The *band-gap thermal sensor* largely overcomes this difficulty. It is a small integrated package which is designed to minimise the effect of self-heating when it draws current from its power supply. Briefly, it does this by having a part of the circuit whose voltage output *decreases* as the current rises, and another part of the circuit whose voltage output simultaneously *increases*. By carefully connecting those parts via resistors, the two effects

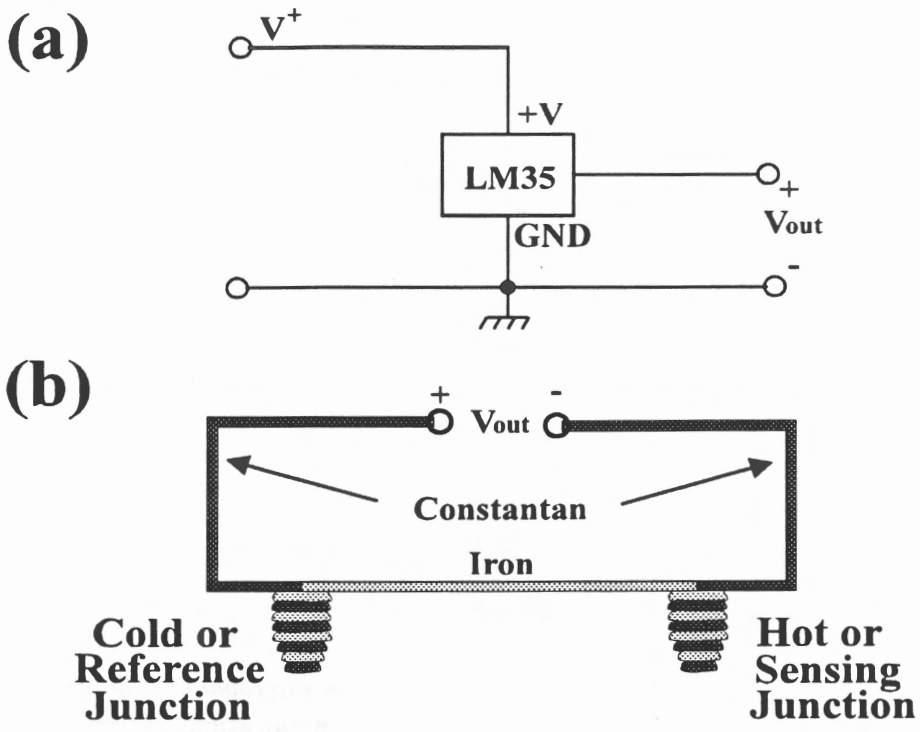


Figure 18.3: (a) A Band-Gap Thermal Sensor (b) A Thermocouple

can be made to cancel, giving a precision output that is independent of the current drawn by the device. You are very likely to encounter a well-known sensor of this type, called the *LM35 precision temperature sensor*, which is sketched in Figure 18.3a. This versatile and interesting device works in the range 0-100°C and has the useful feature that it outputs 0V at 0°C and 1V at 100°C. It is linear to a good degree of accuracy, and so it will work well as a thermometer, even if you simply connect it to a multimeter to display the output voltage.

The last important type of thermal sensor is the *thermocouple*, which is shown in Figure 18.3b. Thermocouples exploit the *Seebeck effect*, which is that a small voltage is produced if you heat the junction of two dissimilar metals. Popular materials for this purpose are iron and a copper-nickel alloy called *constantan*. A simple thermocouple incorporates two junctions (the "cold" junction and the "hot" junction) as you see in Figure 18.3b. It should be clear that, if both of these junctions are at the same temperature,

then the voltage drop at the cold junction will exactly equal the voltage step at the hot junction, and so (by KVL) the output voltage will be zero. If, however, the cold junction is kept at a *reference temperature* while the hot junction is placed in the region whose temperature is to be measured, then the voltage drop and the voltage step will be of different sizes, and (again by KVL) the difference will appear as the output voltage. The output voltage is, in fact, proportional to the *difference* between the junction temperatures, so it is quite simple to calculate the hot junction temperature if you know the reference temperature and the constant of proportionality. Thermocouples are quite linear and the junctions can be made very small. They are convenient devices, and certain specialised junctions are capable of measuring temperatures of over 1000°C. To get a higher voltage output, several thermocouples can be connected in series, with their corresponding junctions grouped together at the hot and cold (or reference) temperatures. Such an arrangement is known as a *thermopile.*

18.4 Sound Sensors

Sound sensors are more commonly-known as *microphones*, and they have an enormous range of applications in telecommunications systems, intruder alarms, vibration detectors and recording apparatus. A very simple, cheap and robust microphone, often found in early telephones, is the *carbon granule microphone*, which consists of a container loosely filled with carbon grains, as shown in Figure 18.4a. A voltage is applied between a flexible metal *diaphragm* at the front of the container and a metal plate at its rear. Current therefore flows through the container via the carbon inside it. If, however, any sound or vibration strikes the diaphragm, it will then apply pressure on the carbon granules, and this has the effect of changing their resistance. The result, by Ohm's Law, is a change in the output current, which closely tracks any pressure variations, and this provides quite an adequate pickup for sound. The output current could then be converted to a voltage, amplified, telegraphed, sent to a loudspeaker, stored etc.

Another extremely effective family of microphones exploits what is known as the *piezo-electric effect.* A piezo-electric crystal is a small piece of quartz or ceramic that has the property that any pressure or vibration applied to its lattice structure will produce a small voltage across it. The *piezo-electric microphone* shown in Figure 18.4b transmits sound energy to a piece of such crystal through the vibrations of a diaphragm, and it then takes the result-

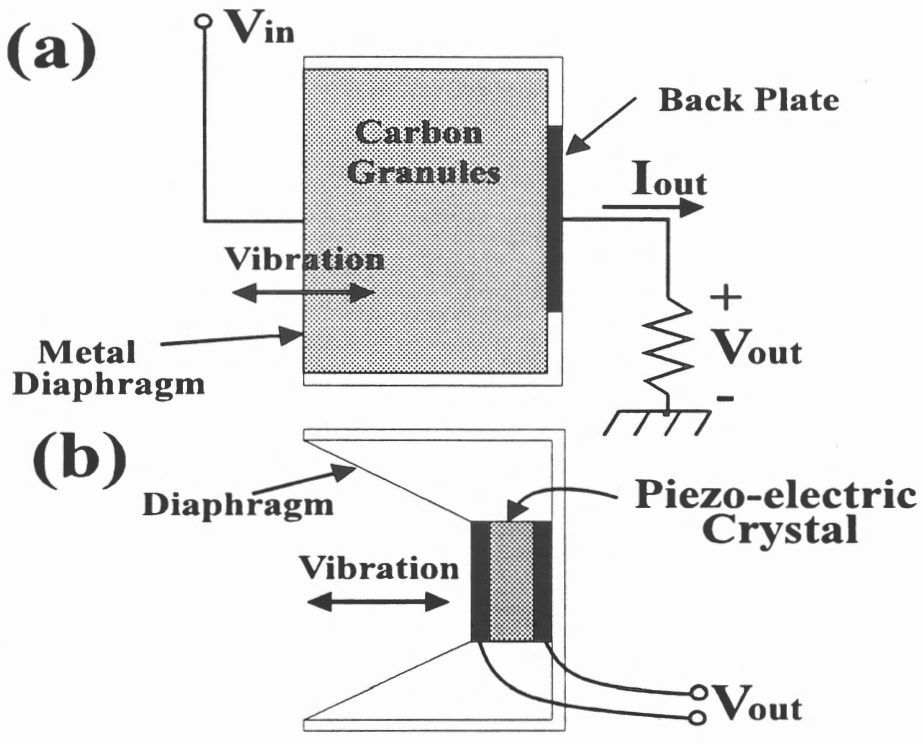


Figure 18.4: (a) Carbon Granule Microphone (b) Piezo-Electric Microphone

ing voltage changes as output. Unfortunately, the output resistance of such microphones is rather high, so the signal has to be amplified using special amplifiers with extremely high input resistance (this means using FETs in their front end). Furthermore, the amplifier requires its own power supply (although the crystal itself does not) and this can be a little bulky. In spite of these practical issues, piezo-electric microphones are extremely effective sound sensors, and they are finding valuable uses in work under water and at the high frequencies known as *ultrasound*. Modern medical scanners and oceanographic research are just two of the applications to which they are being put.

Perhaps the best-known microphone is the *electromagnetic microphone*, in which a diaphragm vibrates with incident sound, causing a small coil, attached to the diaphragm, to move. The coil sits within a magnetic field, provided either by a permanent magnet or else by a field coil separately supplied with dc. Movements of the small coil within this field induce a

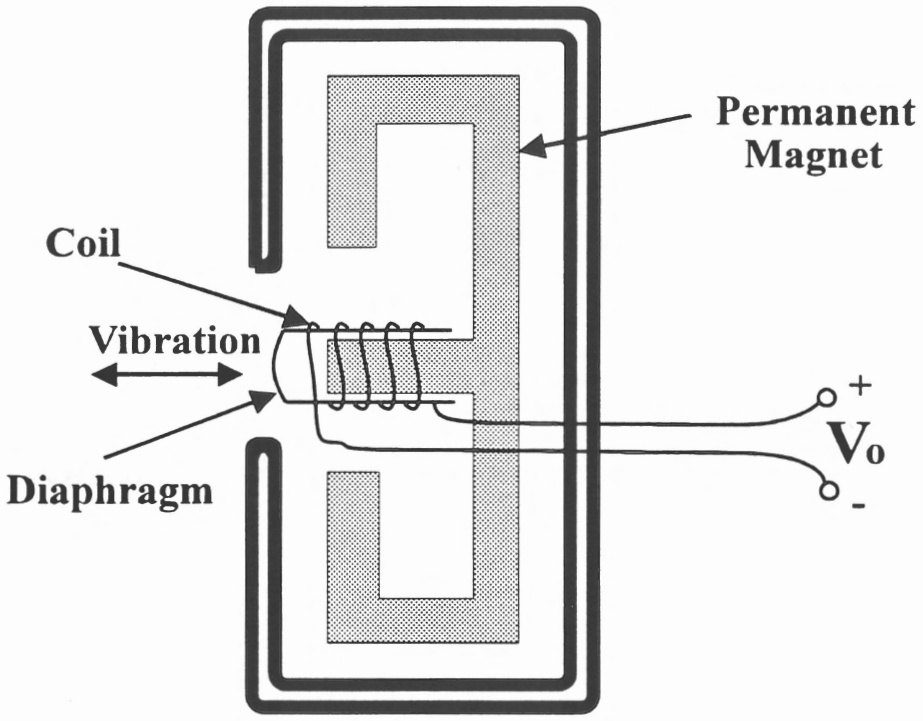


Figure 18.5: A Moving-Coil Electromagnetic Microphone

fluctuating voltage, which can be taken off as output. Since the voltage is proportional to the movements of the coil, it represents a record of all of the diaphragm movements, and as such it gives a faithful conversion of the sound into electrical form. There are certain advanced practical difficulties with these devices (such as the tendency of the small coil to resonate), but the moving-coil electromagnetic microphone is a compact and linear device with an enormous range of uses. A sketch of the microphone is given in Figure 18.5.

Chapter 19

Sensors & Transducers (2)

19.1 Magnetic Sensors

We continue our brief look at sensors and transducers by mentioning a few devices that respond to *magnetic fields.* These not only support modern instruments for direction finding and orientation in the earth's rather weak magnetic field, but have many other applications in measuring the magnetic fields that arise in electro-mechanical, medical and navigational studies. We have already seen, in the case of the electromagnetic microphone, an *inductive* magnetic sensor in the shape of the moving coil. The emphasis here will therefore be on some other types of magnetic sensor that do not rely on the inductive effect.

We begin with the *Hall effect sensor.* Currents that flow in semiconductors are explained, as we know, by the motion of charged carriers liberated from the atoms of the lattice. If a magnetic field is applied to a semiconductor in which a current is flowing, then the conduction electrons will be attracted by the field to one side of the semiconductor in preference to the other. The direction in which the electrons tend to move can be found conveniently by *Fleming's Left Hand Rule*, which it is assumed that you know well. The result is shown in Figure 19.1a, which illustrates how a *voltage* appears across the semiconductor slice because there are now more electrons on one side than on the other. Although this voltage is small, it can be amplified and measured, and so the Hall effect device gives us a sensitive (and, in certain Hall effect devices, linear) way to detect the presence, quantify the strength or deduce the direction of an applied magnetic field. Some simple Hall effect sensors are designed to operate as on-off switches, only changing their state

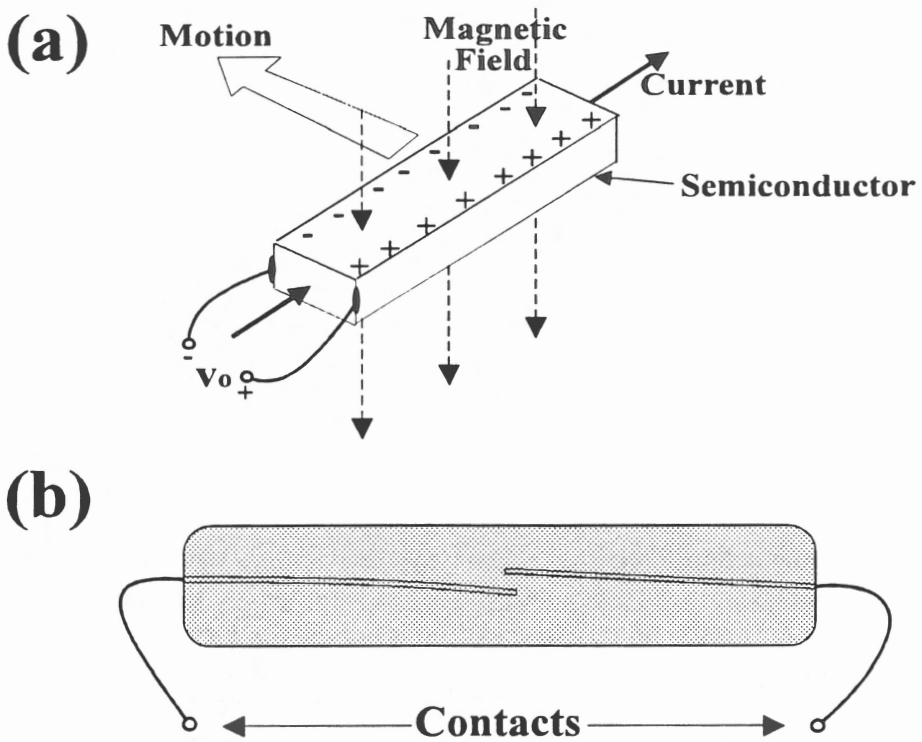


Figure 19.1: (a) The Hall Effect (b) A Reed Switch

if the magnetic field exceeds a certain strength or has a certain direction. These are known as *Hall effect switches*, and they are useful in determining the proximity of a permanent magnet. You can see them used, for instance, in magnetic burglar alarm fixtures that sound the alarm when a window is opened, thereby separating a small magnet from a Hall effect sensor.

There are certain quite exotic metallic alloys whose resistance varies with the strength of the local magnetic field. These are called *magnetoresistive sensors*. If a steady current is supplied to a resistor made from such a material, then the voltage across the resistor will change even if there are small changes in local magnetic field. A magnetic storage tape being "swiped" past the resistor is enough to cause fluctuations in the output voltage across the resistor, and these fluctuations can be *interpreted* so as to read the data stored on the tape. The recent explosion in the use of bank cards, for example, owes much to this technology.

Figure 19.1b shows a *reed switch*, in which two sprung metal contacts are sealed in a capsule. When the switch is within a magnetic field, the two contacts become temporarily magnetised, and will attract each other rather like pins or iron filings do when a magnet is brought close to them. Upon touching, they close a circuit. Reed switches are also often found in magnetic burglar alarm switches on doors and windows, allowing the open door or window to be detected by the security system. Hopefully you can see clearly how the output voltage from a suitably-connected reed switch could be used in logical circuitry as part of a more complex alarm circuit. One might design a system, for example, that sounded an alarm if a door were opened OR a light went on OR a sound was heard in a house. The three inputs to the logical circuitry might come from a reed switch, a phototransistor and a piezo-electric microphone respectively, and would be connected via a 3-input OR gate to an *audible warning device* such as a siren or bell.

19.2 Mechanical Force Sensors

Great engineering ingenuity has been devoted over many years to the production of devices that can register or measure *forces* of various kinds. If you ever want to test how a structure or a substance performs, or is likely to perform, under real conditions over time, then it is vital to be able to apply carefully-quantified forces to it. You would also want to measure the effects of the applied forces, even though they may be invisible (unless destruction ensues!). Aside from this, *pressure* and *weight* are two forces which we often need to measure accurately in our scientific or commercial endeavours. We will now take a very quick look at how certain types of force measurement are achieved, and at some of the simplest sensors that are designed to do the job.

Figure 19.2a shows a *metal foil strain gauge*, and two of its many applications are illustrated in Figure 19.2b and Figure 19.2c. As you see, the strain gauge consists of an extremely thin metal conductor, laid mostly along parallel lines, and attached to the surface of an object to which forces are to be applied. The resistance of the conductor is given by a formula that we saw back in Chapter 1:

$$R = \frac{\rho L}{A} \tag{19.1}$$

If the gauge is subjected to any stretching or *tension*, then its length, L, increases slightly whilst its cross-sectional area, A, decreases. Hence, its

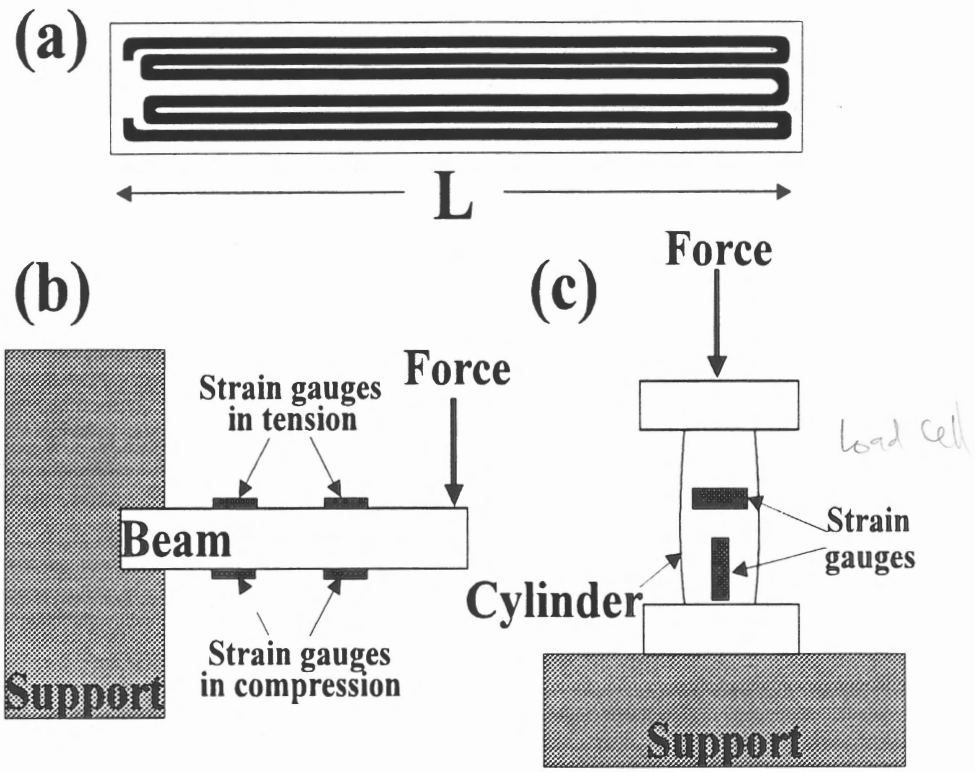


Figure 19.2: A Strain Gauge and Two Applications

resistance will rise. Conversely, any *compression* of the gauge will reduce its length and slightly thicken its cross-section, leading to a drop in its resistance. In a well-designed strain gauge, the resistance responds *linearly* to the force applied. The percentage change in resistance over the percentage change in length is called the *gauge factor*, and is typically about 5 for a metal foil strain gauge. With a steady current supply and suitable amplification of the output voltage across the gauge, therefore, a sensitive response to very small changes in the shape of the gauge (or whatever it is attached to) may be expected. Figure 19.2b shows how strain gauges may be positioned to determine the effects of loading a thick beam, while Figure 19.2c shows a similar analysis being carried out on a metal shaft which is to be loaded along its axis by a weight.

A good deal of design goes into ensuring that the temperature of the strain gauge does not affect its performance, but these interesting details must be

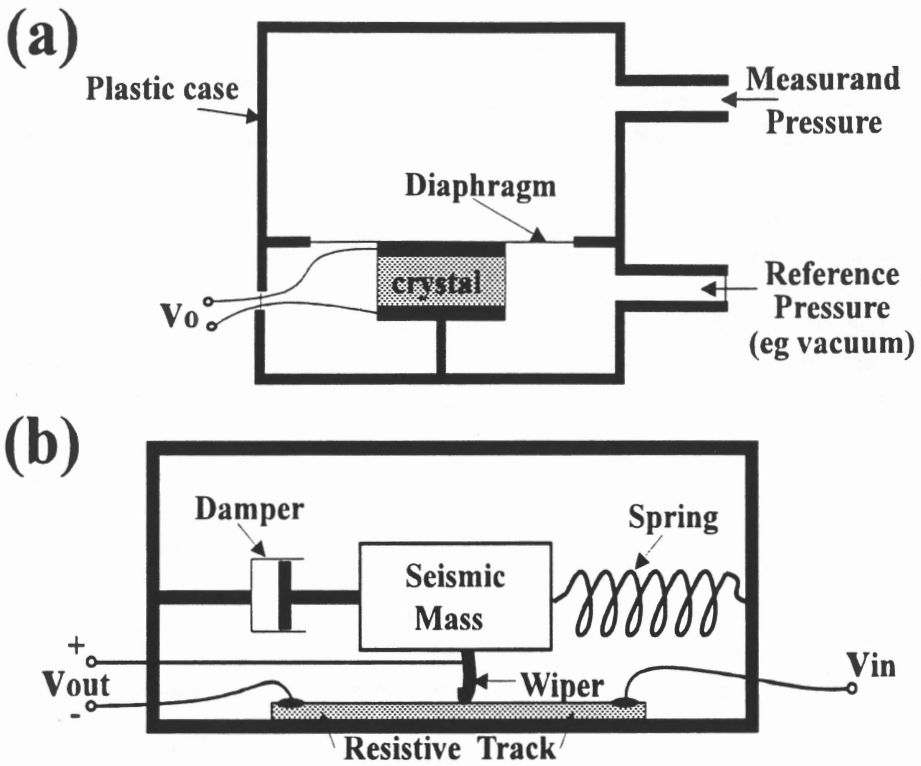


Figure 19.3: (a) A Pressure Sensor (b) An Accelerometer

left for later courses. For the present, you should note that Figure 19.2c has an obvious application: we could use the metal cylinder as a force-sensing element in its own right (known as a *load cell*). This would mean that instead of using the strain gauge to assess the effects of applying the force, we are using the metal cylinder to *measure* the force applied. The linearity of the strain gauge and of the load cell's response to applied weight, therefore, would give us the basis for an *electronic weighing scale*. The output voltage would be taken from the strain gauge and suitably amplified and displayed to give a reading of the weight sitting on the cylinder.

Figure 19.3a presents a piezo-electric pressure sensor that might be of value in an aircraft, submarine or weather station. The amount of force on the piezo-electric crystal varies with the pressure in the upper chamber, which pushes down on a diaphragm attached to the crystal. The output voltage is then amplified and displayed. You should see that the crystal could be re-

placed by a small solid block to which a strain gauge was attached, and one would still have a pressure sensor if the output from the strain gauge were then amplified and displayed. The piezo-electric device is a true *transducer* while the strain gauge idea is not.

The last mechanical force which we consider measuring is the *accelerative* force, which is of great interest in many areas including testing vehicles and machinery or studying mechanical vibrations or shock (e.g. the effects of minor earthquakes on buildings). As we know from basic physics, acceleration is the result of applying a force to a *mass*, and consequently the best way to measure an accelerative force is to sense its effects upon a mass within our measuring instrument. Figure 19.3b shows an *accelerometer* which contains a small *seismic mass* that will move with respect to the rest of the instrument if the whole device is accelerated to the left or to the right. The distance that the mass moves can be registered by the change in voltage between the two output wires shown, since the mass is attached to a *wiper* that runs along a resistive track. The greater the acceleration of the whole instrument (or of whatever it is attached to), the greater the movement of the mass from its rest position. The output voltage therefore represents the accelerative force acting upon the sensor (or upon whatever it is connected to). The *spring* is there to restore the seismic mass to its rest (or zero) position when there is no accelerative force acting, and the *damper* serves to minimise unwanted oscillations of the seismic mass from side to side. Very small accelerometers are available (measuring a few millimeters on each side) and these can inform us about the acceleration of much larger objects without undue intrusion. Piezo-electric accelerometers also exist, and you may like to imagine how such a device might work.

19.3 Displacement Sensors

The last type of sensing device that we will consider in this course is the *position* or *displacement sensor*. As you will appreciate, we have only been able to look at a fraction of all of the sensor types available. It is left to your later studies to investigate humidity or acidity sensors, flow meters, capacitance meters etc., and to go more deeply into the ideas sketched out in this course. Similarly, we will restrict ourselves here to just *three* of the many available methods of measuring the precise position of an object. You will doubtless meet up with interesting improvements on these in later courses.

The very simplest position sensor is the humble potentiometer! A linear

potentiometer is just a variable resistor whose *wiper* slides along a resistive track, varying the resistance between itself and the end of the track in proportion to its distance from the end of the track. Many of the circular *rotary "pots"* which you have seen in the laboratory are also quite linear over most of their range, but beware - some are not! If an object is connected to the wiper, then the output voltage between the wiper and the end of the track is proportional to the position of the object, assuming that the voltage across the whole "pot" stays constant. This is, of course, the principle that we saw in our simple accelerometer of the previous section. The advantages of this method are that it is very simple and easy to implement; also, the output voltage can be made large just by using a suitably large voltage across the whole potentiometer, so amplification may not be necessary. The disadvantage is that there is friction and wear between wiper and track, which can degrade the mechanical performance and eventually lead to a loss of electrical linearity. Note that the use of a *rotary* potentiometer offers possibilities for sensing the *angle* of any object attached to its central spindle. The main point to grasp about either of these sensors is that it is best if the output voltage is proportional to the measurand.

The *linear variable differential transformer* (or *LVDT*) shown in Figure 19.4a offers a completely different approach to displacement sensing. An alternating current is passed through the primary coil of a transformer whose straight core is free to move. Two coils, wired in series with each other but wound in opposite senses, form the secondary circuit. When the core is in its central position, equal voltages are *induced* in both of the secondary coils, so (by KVL) the output voltage is zero. However, if the core is displaced to either side, it then permits a greater magnetic linkage of the primary with whichever of the two secondary coils is wound around more of it. That coil will then have a greater voltage induced in it, and so the output voltage (again by KVL) will rise in amplitude by an amount that is related to how far the core has been displaced. If an object is now attached to the core, then the LVDT provides a method of returning quite precise information about the object's position, via the output voltage. Note, again, that if the transformer core used were not straight, but formed part of a circle, then you could design a sensor which measured the *angular displacement* of the attached object.

Fianlly, let us look at the *optical position sensor*, one version of which is shown in Figure 19.4b, to get the flavour of a third and again different method of displacement sensing. Optical position sensors will include one

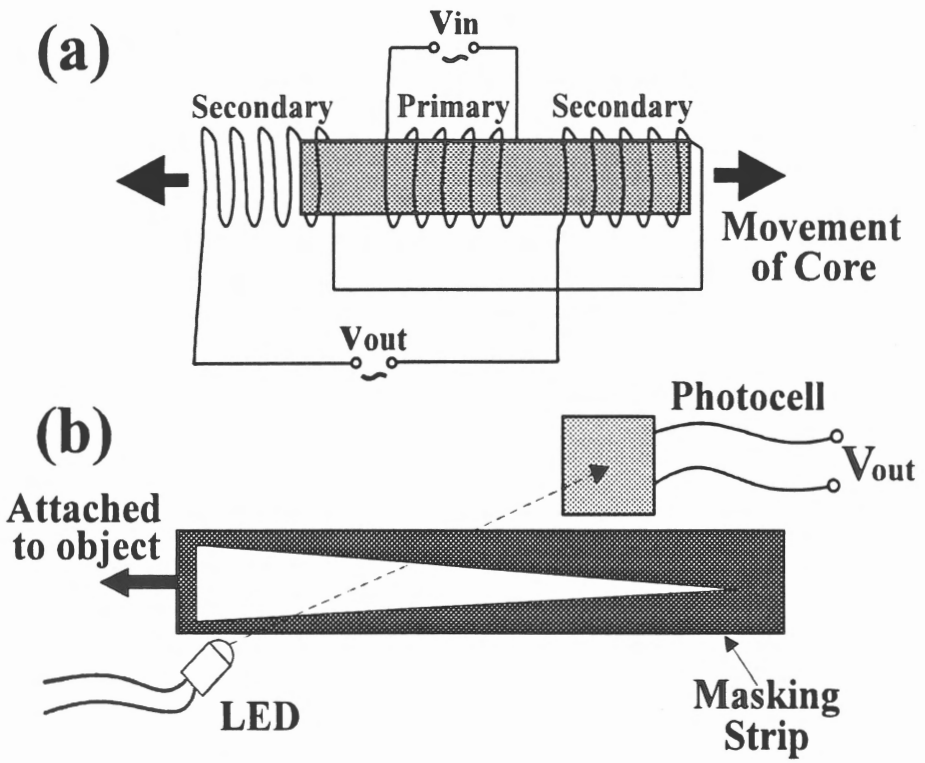


Figure 19.4: (a) An LVDT (b) An Optical Position Sensor

or more visible or infra-red light sources (such as LEDs) and also one or more sensors for this radiation (typically photodiodes). The object whose position is to be measured is connected either to the sources or to the sensors in such a way that, as it moves, the output from the sensors changes. This can be achieved by having a number of photodiode sensors, each one of which represents a certain position of the object if it "sees" the LED. More economically, the moving object can be attached to a strip with one or more transparent windows in it. Movement of the strip can then be interpreted by the sensing circuitry to yield a measure of the strip's position, and hence that of the object. There are many very interesting variants on this sort of idea, and you can research these for yourself. The position sensor shown in Figure 19.4b is a delightfully simple version that requires only one LED and one radiation device, a photovoltaic cell. If the object is positioned far to the left, then the LED is entirely hidden from the photovoltaic cell by the opaque part of the strip. A triangular window cut into the strip ensures,

however, that as the object moves to the right, a greater and greater fraction of the cell will be illuminated. Moreover, the percentage illumination increases linearly with the object's motion to the right, because the triangular window has straight-line sides. Since the fraction of the cell that is illuminated is directly proportional to the output voltage of the cell (we saw this in the previous chapter), it follows that the output voltage is a linear representation of the object's position. Many elegant designs of this sort are available, and they can be adapted to give angular information or even velocity measurements for the object. This, however, is the simplest, and it is recommended to your attention for a start.

With that we conclude our brief coverage of the world of sensors. It should now be clear where many of the "signals" that we have thought of amplifying in previous chapters actually originate. Hopefully, this provides a better focus for why logical circuits, switches and amplifiers are worthy of study, as well as giving an introductory insight into a major and fascinating branch of electrical engineering.

Chapter 20

Integrators & Differentiators

20.1 Amplifiers With Impedances

You will now be familiar with the fact that an inverting op-amp-embedded amplifier circuit with input resistance R_a and feedback resistance R_f will yield an output voltage of

$$v_o = -\frac{R_f}{R_a} v_i \tag{20.1}$$

where v_i is the input voltage. Your work on impedances in the other half of this course should convince you that *any* impedances in the input or feedback branches of the op-amp circuit (resistors, capacitors or inductors) will give a similar result. Thus, in Figure 20.1a

$$v_o = -\frac{Z_f}{Z_a} v_i \tag{20.2}$$

All of the analysis methods that we have so far developed for ideal op-amp circuits (including the three main Rules) can be applied with capacitors or inductors present. From your work with electrical circuits, you will hopefully find it natural to replace a capacitance, C, with an *impedance* of $Z_C = 1/(j\omega C)$, and an inductance, L, with $Z_L = j\omega L$, where ω is the *frequency* of the input. You should also remind yourself of the following relationships:

$$i_c = C\frac{dv_c}{dt} \qquad v_c = \frac{1}{C}\int_0^t i_c\,dt + V_o \qquad i_L = \frac{1}{L}\int_0^t v_L\,dt + I_o \qquad v_L = L\frac{di_L}{dt} \tag{20.3}$$

By applying our ideal op-amp Rules and simple nodal analysis, you will find it easy to analyse op-amp circuits that include impedances.

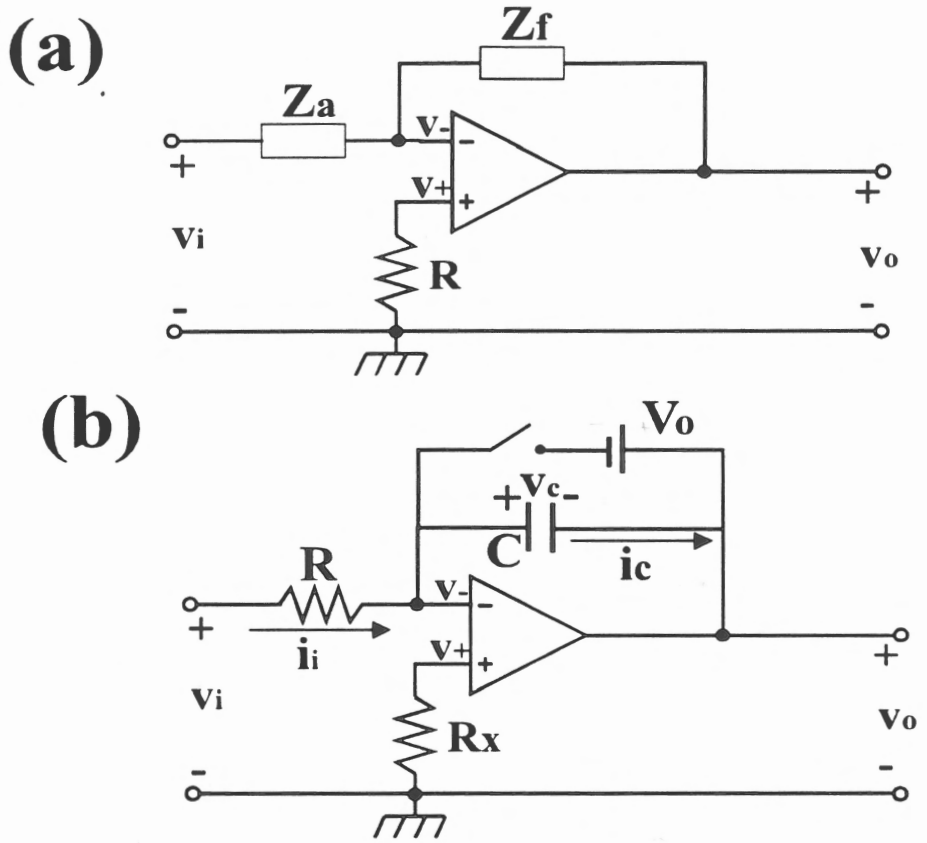


Figure 20.1: (a) Impedances in an Amplifier (b) An Inverting Integrator

To illustrate the point, we will now look at two very important op-amp circuits. The first of these is given as Figure 20.1b, in which the feedback resistor of a conventional inverting amplifier has been replaced by a *feedback capacitor*. The circuit has an *ac* input voltage, v_i, that may vary periodically or even at random. We quickly see that

$$v_- = v_+ = 0 \tag{20.4}$$

from which we know that a current of

$$i_i = \frac{v_i - v_-}{R} = \frac{v_i}{R} \tag{20.5}$$

is flowing in the input branch of the circuit. Because of the op-amp's very high input resistance, almost all of this current (remember it is ac) passes

through the capacitor, provided that the switch is open. Thus,

$$i_c = i_i = \frac{v_i}{R} \tag{20.6}$$

If the switch were closed, then all of the current would pass through the switch and battery, and the capacitor voltage would be V_o. With it open, since the capacitor voltage as shown in Figure 20.1b is, in general,

$$v_c = \frac{1}{C}\int_0^t i_c\,dt - V_o = \frac{1}{RC}\int_0^t v_i\,dt - V_o \tag{20.7}$$

we see that the output voltage of the circuit is

$$v_o = 0 - v_c = -\frac{1}{RC}\int_0^t v_i\,dt + V_o \tag{20.8}$$

What this equation tells us is that if, at time $t = 0$, the switch is opened, then current will flow in the capacitor which will begin to discharge (if v_i is positive) from its initial voltage of V_o. In fact, the capacitor voltage is proportional to the (negative) integral of the input voltage, so the circuit of Figure 20.1b is called an *inverting integrator*. Check for yourself that if $V_o = 0$, $R = 1\text{k}\Omega$ and $C = 1\mu\text{F}$, and an ac input voltage of $v_i = 2\cos(50t)\text{mV}$ is applied, then the output voltage will be $40\sin(50t)\text{mV}$. The amplification factor of 20 could be increased by reducing the value of either R or C.

The inverting integrator is an extremely useful and important circuit that you will study very thoroughly in later courses. For the moment, the main aim should be to make sure that you understand *why* it performs as an integrator, and to convince yourself that this behaviour can be deduced from nothing more than our three ideal op-amp Rules.

Figure 20.2a shows a similar op-amp circuit, but this time a capacitor is placed in the *input* branch, and there is a resistor in the feedback loop. Once again, the input voltage is assumed to be ac (because the input capacitor could not pass a dc signal). Analysis now gives us $v_- = v_+ = 0$ and so $i_i = C\,dv_i/dt = i_f$. Hence

$$v_o = -Ri_f = -RC\frac{dv_i}{dt} \tag{20.9}$$

Evidently, we have an *inverting differentiator*. We ought to stress here that there are practical difficulties with this circuit that prevent it from working as well as the theory suggests - but we will ignore these for now. You might

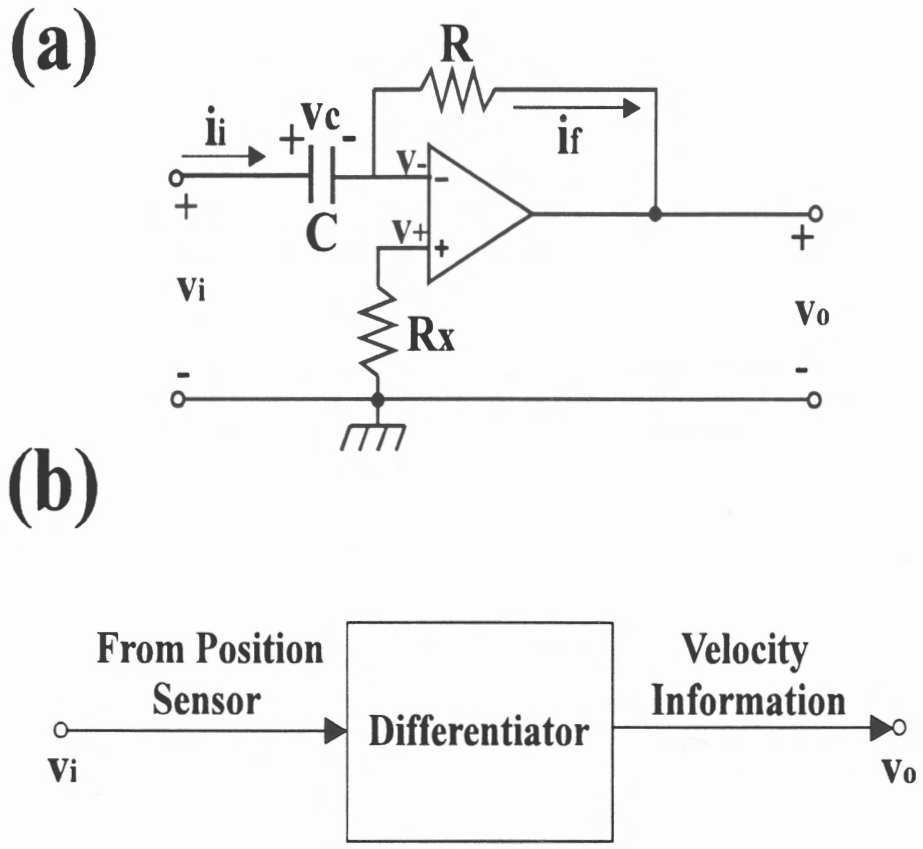


Figure 20.2: The Inverting Differentiator

like to consider the value of a circuit like the one shown in Figure 20.2b, if v_i were the output from a displacement-measuring device such as the ones we saw in the previous chapter. The output of the differentiator, v_o, would then represent the *rate of change* of displacement, or *velocity*, of the object attached to the sensor. Likewise, you can visualise circuits whose outputs give the rate of change of light intensity, rate of change of temperature, rate of change of applied force, and so on. In principle, differentiator circuits open up a whole new range of physical quantities that we can measure in the world around us, and they offer us a suite of new techniques for electronic measurement. You will need the theory given in some later courses, however, before you can judge their advantages and disadvantages in practice.

20.2 Analog Computers

You are, however, ready now for a glimpse of one of the most fascinating applications for op-amps. When we first introduced op-amps, we pointed out that the word "operational" in their name arises from their ability to perform mathematical *operations*. By now, we have seen their rôle in circuits that add, subtract, multiply, divide, integrate and differentiate the input voltage or voltages, and this suggests that we could build quite simple circuits that would perform a *sequence* of these operations. As the input or inputs to such circuits were varied, the output or outputs would represent *solutions* to mathematical equations. In particular, since we now have integrators and differentiators, we could get outputs that are solutions to *differential equations*. When we use a circuit in this way, we call it an *analog computer*, and you ought to be aware that analog computers are widely used, and that they solve equations very much *faster* than the *digital computers* with which we are all familiar.

As an example, let us look at an analog computer that will *solve* the differential equation governing the circuit of Figure 20.3a. You will quickly see (from your work on electrical circuits) that applying KVL to this circuit gives

$$v(t) = L\frac{di}{dt} + \frac{1}{C}\int_0^t i\,dt + Ri \tag{20.10}$$

assuming that the capacitor voltage is initially zero. Now, to solve for the current $i(t)$, we note that this equation gives

$$\frac{di}{dt} = \frac{v(t)}{L} - \frac{Ri}{L} - \frac{1}{LC}\int_0^t i(t)\,dt \tag{20.11}$$

or, equivalently,

$$i(t) = \int_0^t \left(\frac{v(t)}{L} - \frac{Ri(t)}{L}\right) dt - \frac{1}{LC}\int_0^t \int_0^t i(t)\,dt\,dt \tag{20.12}$$

Now look at how the circuit of Figure 20.3b represents these equations. You should start by noting that the *output* (a voltage which represents the current i(t)) is to be taken at the output of the second op-amp, as shown. This choice is indicated by the fact that $i(t)$ appears both integrated once and also integrated twice on the right hand side of the equation above. The third op-amp is part of an inverting integrator whose resistor, R_e, is chosen

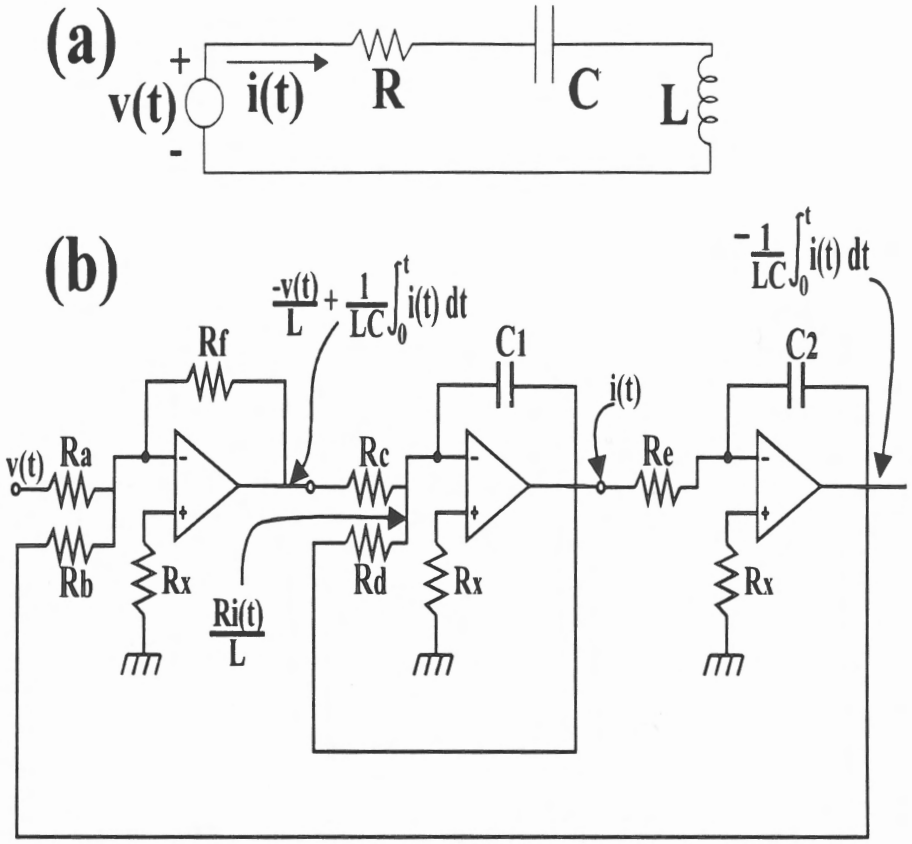


Figure 20.3: A Simple Analog Computer to Solve a Differential Equation

so that the op-amp output will be

$$v_{o3} = -\frac{1}{LC}\int_0^t i(t)\,dt \tag{20.13}$$

This output is *fed back* to the first op-amp, where it is put through an inverting summer, together with the input voltage, $v(t)$, to form

$$v_{o1} = -\frac{v(t)}{L} + \frac{1}{LC}\int_0^t i(t)\,dt \tag{20.14}$$

Again, R_a, R_b and R_f are chosen so as to provide the gain of $1/L$. Lastly, this expression is summed with a fraction of $i(t)$, set by R_d to be equal to $Ri(t)/L$, and then inverted and integrated in the second op-amp section, so

that the output of the second op-amp truly is

$$v_{o2} = \int_0^t \left(\frac{v(t)}{L} - \frac{Ri}{L} \right) dt - \frac{1}{LC} \int_0^t \int_0^t i(t)\, dt\, dt \qquad (20.15)$$

Now, you may well ask what the point of this is, if all we are going to do is to watch $i(t)$ vary in the analog computer while the input $v(t)$ changes as it did in the original circuit. Why not simply measure $i(t)$ in the original circuit with an ammeter, and save a lot of trouble?

We have to remember that the original circuit of Figure 20.3a may represent a very large or complex circuit in real life. The current, $i(t)$, may be inaccessible if, for example, the place where it flows is tremendously hot or noisy, or if it is a long way away (such as on a probe orbiting another planet). Alternatively, the circuit of Figure 20.3a may not yet be built, but we want to *model* how it would perform before we go to the trouble and expense of building it. There are several more sophisticated reasons (which you will study in courses in *control theory*) for why we might want to create a small circuit that mimics a much larger one, and this is where analog computers are of great value. Although at first the method of designing the analog computer circuits may seem difficult, you should study the example given in this section very carefully, and try to see how you would put together circuits to solve other simple equations, differential or otherwise. In later courses, you will see some amazing possibilities that arise from this type of electronic circuit.

Chapter 21

Further Op-Amp Applications

One of the problems of bringing a short book such as this one to a close is that there are no very neat endings in a field as wide as electronics. We have come a long way to get from a description of the silicon atom right up to a simple analog computer made of op-amps. In the process, however, we have always had to *simplify* for the sake of the clarity that is desirable in an introductory course. The penalty that one must pay for this is that the end of the story seems rather ragged, because electronics is a rich and convoluted subject which one should not attempt to simplify too far. The reader is therefore asked to accept a final chapter which merely lists a few further op-amp circuits. These are circuits which illustrate some of the many interesting ways in which op-amps are used, and which in some cases point the way forward to more advanced electronic concepts. The list is not remotely exhaustive, however, for such an undertaking would double the length of this course! Instead, it is hoped that you will use the circuits that follow for further examples of op-amp circuit analysis, and that you will be inspired, by some of the new ideas that these circuits bring, to explore the world of electronics at a higher level.

21.1 The Optional Inverter

The first circuit that we will mention is shown in Figure 21.1. Here, an op-amp connected with a feedback loop has its inverting terminal supplied by an input voltage signal, v_i, while the non-inverting terminal input may

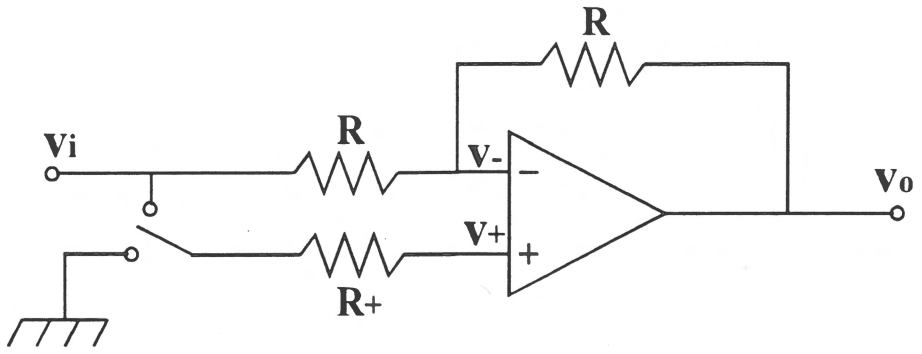


Figure 21.1: An Optional Inverter Circuit

be switched between v_i and ground. Let us analyse the circuit with the switch in each position. Firstly, if the switch connects the resistor R_+ to ground, then we have a virtual earth at the v_- terminal by op-amp action. A current of v_i/R flows into the input and through the feedback loop, so it is easy to see that

$$v_o = 0 - \frac{v_i}{R}.R = -v_i \tag{21.1}$$

In other words, we have a *unity gain inverter.* With the switch in the other position, connecting R_+ to v_i, both op-amp terminals will be forced to equal v_i. Then, since no current flows from the input into the feedback loop, the output voltage simply equals the input voltage, and we have a *unity gain follower* with

$$v_o = v_i \tag{21.2}$$

You can see that the position of the switch determines whether the output voltage is an inverted or a non-inverted copy of the input voltage. For this reason, the circuit is called an *optional inverter.*

We cannot go here into the value of circuits such as this, except to mention that they are of great importance in the production of what are called *amplitude modulated* or *AM* signals. Imagine replacing the mechanical switch, shown in Figure 21.1, by an electrical switch such as a BJT or FET. You could then switch extremely rapidly and often between inverting and not-inverting an input signal. The output from the optional inverter circuit would look rather messy on a graph, but it has certain properties that are desirable when it comes to broadcasting the signal for considerable distances across the air.

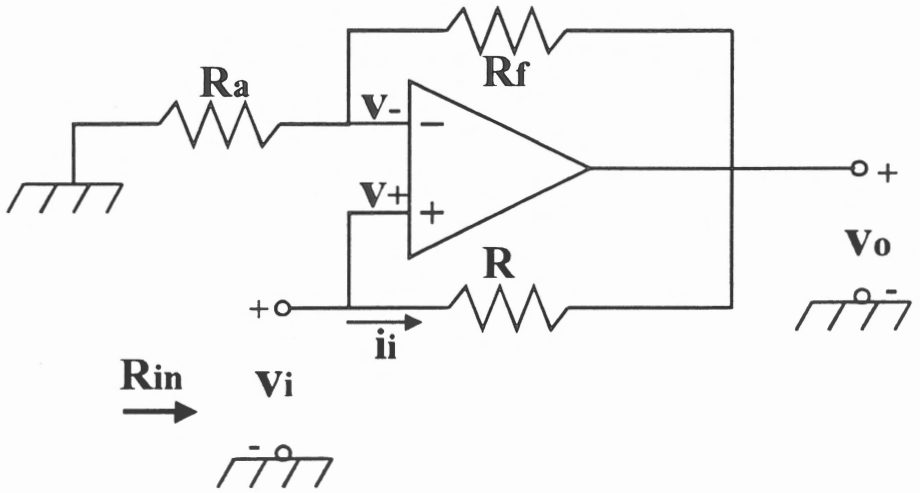


Figure 21.2: A Negative Resistance Circuit or NRC

21.2 The Negative Impedance Circuit

One version of another very interesting circuit is shown in Figure 21.2, in which an op-amp is connected with a negative feedback loop and has its inverting terminal grounded. The non-inverting terminal is also connected to the output via a resistor R. Input to the circuit is into the non-inverting terminal, as shown. Let us consider the *input resistance* of the circuit. Clearly, this is

$$R_{in} = \frac{v_i}{i_i} \tag{21.3}$$

Now, by voltage division we can see that

$$v_- = \frac{R_a}{R_a + R_f} v_o \tag{21.4}$$

and, since $v_- = v_+ = v_i$, we can write this with v_o as the subject:

$$v_o = v_i \left(1 + \frac{R_f}{R_a}\right) \tag{21.5}$$

Virtually all of the current i_i passes through R because of the op-amp's huge input resistance, and so we can use Ohm's Law to write

$$i_i = \frac{v_i - v_o}{R} = \frac{v_i - v_i(1 + R_f/R_a)}{R} = -\frac{R_f v_i}{R_a R} \tag{21.6}$$

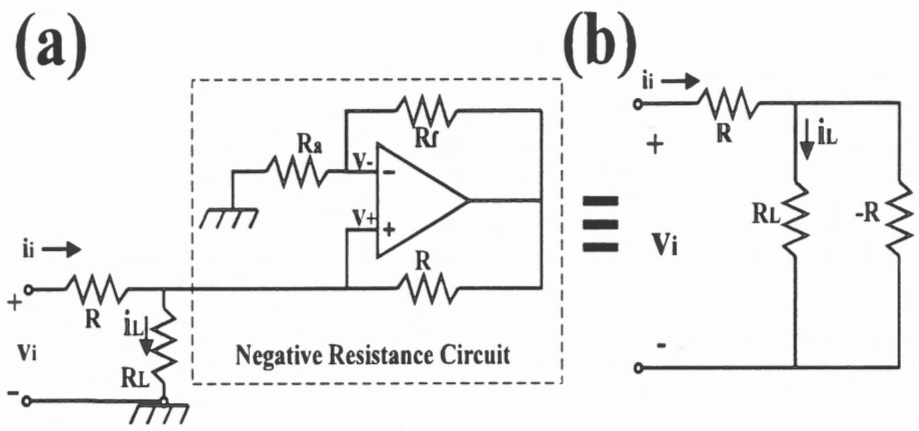


Figure 21.3: A Dependent Current Source Made Using a NRC

The input resistance of our circuit is therefore given by

$$R_{in} = \frac{v_i}{i_i} = -\frac{R_a}{R_f} R \tag{21.7}$$

We are forced to the conclusion that the circuit acts as a *negative resistor.* If $R_a = R_f$, then the size of this negative resistance is exactly $-R$. An identical development will show that replacing R with any impedance, Z, will yield a *negative impedance* of value $-R_a Z/R_f$, looking into the input terminals of this circuit. The circuit is therefore called a *negative impedance circuit* or *NIC.* Our example is, of course, a *negative resistance circuit.*

This is nice mathematics, but what does it mean? It seems rather difficult to imagine what a negative resistance would be like, let alone if it has any uses! Ohm's Law

$$i = \frac{v}{R} \tag{21.8}$$

would suggest that, for $R < 0$, the higher the voltage across its terminals, the greater the current flowing *out of* the positive terminal of such an element. In many ways, this is like a *battery*, and negative resistors are indeed used in *current generating* circuits, such as the one shown in Figure 21.3a. A load resistor is connected to a NIC so that both are effectively in parallel down to ground. The NIC is designed with $R_a = R_f$, so its resistance $-R$ exactly balances the resistance R in the input branch of the circuit. An equivalent circuit, with the NIC replaced by a resistance $-R$, is shown in Figure 21.3b. We now analyse this circuit to find the load current, i_L.

The input resistance of the whole circuit is

$$R_{in} = R + R_L//(-R) = R - \frac{RR_L}{R_L - R} \tag{21.9}$$

so the input current is given by

$$i_i = \frac{v_i}{R_{in}} = \frac{v_i}{R - RR_L/(R_L - R)} = \frac{v_i(R_L - R)}{RR_L - R^2 - RR_L} = -\frac{v_i(R_L - R)}{R^2} \tag{21.10}$$

It is now an easy matter to find the load current by current division, after which we substitute our expression for i_i, to obtain

$$i_L = \frac{-R}{R_L - R} i_i = \frac{-R}{R_L - R} \cdot \frac{-v_i(R_L - R)}{R^2} = \frac{v_i}{R} \tag{21.11}$$

We come to the remarkable conclusion that the load current is *independent of the size of the load resistor and depends on the input voltage.* Clearly, we have the basis for a *dependent current source*, which would be an extremely useful device if, for instance, the resistance of the load was variable, but we wanted to be sure that it was supplied with a fixed current. Similar useful circuits are possible if the NIC is used to develop a negative capacitance or negative inductance, and you may meet some of these in your future studies.

21.3 V-to-I and I-to-V Convertors

The current-generating circuit of the previous section is only one of many dependent-source circuits that can be made from op-amps. Figure 21.4a presents anpther simple *voltage-to-current convertor* in which the load, R_L, is to be supplied with a stable current controlled by the input voltage, v_i. Check that you agree that, since the op-amp inputs are high-resistance

$$v_+ = v_i \quad \text{and} \quad v_- = i_L R_x \tag{21.12}$$

Then, since $v_+ = v_-$, we easily arrive at

$$i_L = \frac{v_i}{R_x} \tag{21.13}$$

so the load current is independent of R_L and is set by varying v_i. Remember that the output current of the op-amp is limited by the voltage rails supplying the op-amp and by the resistances in the output stage, so the circuit

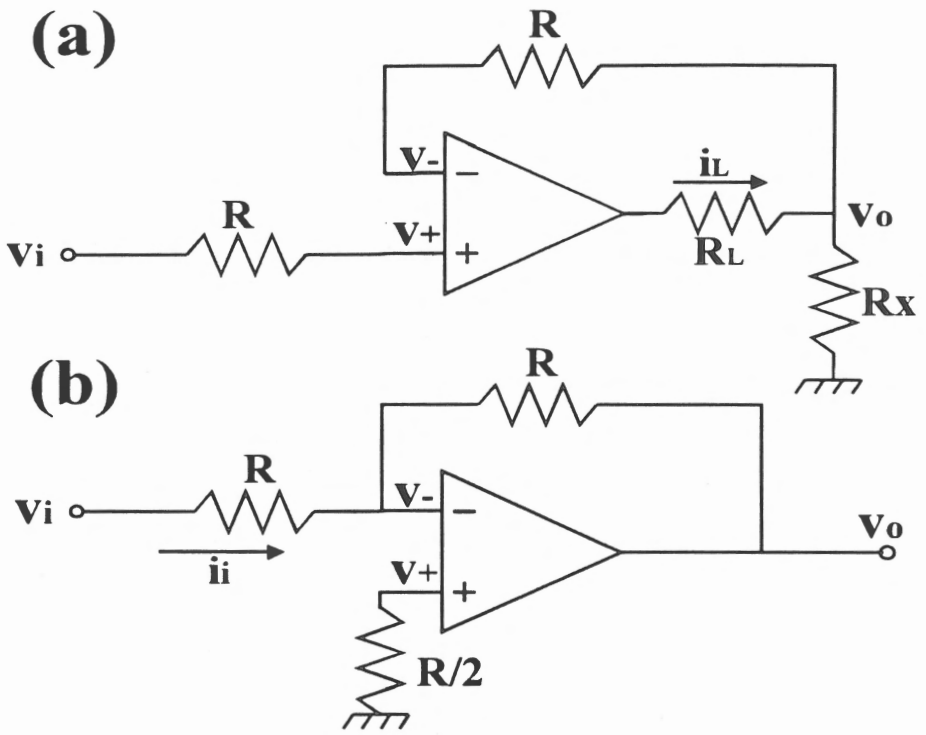


Figure 21.4: (a) V-to-I Convertor (b) I-to-V Convertor

cannot supply limitless current to the load, no matter how v_i is increased. With this limitation, however, the circuit of Figure 21.4a will work well as a dependent current source over a range of current values, commensurate with the op-amp's ability to supply current.

Figure 21.4b shows another simple circuit, which we can regard as a *current-to-voltage convertor*. Again, we can analyse this circuit using the familiar properties of ideal op-amps. The high op-amp input resistance means that $v_- = v_o + i_i R$ by Ohm's Law, and, since $v_- = v_+ = 0$, we see that the output voltage is proportional to the input current:

$$v_o = -i_i R \tag{21.14}$$

This gives some idea of how the dependent voltage sources that we sometimes see in electrical circuit analysis can be achieved in practice.

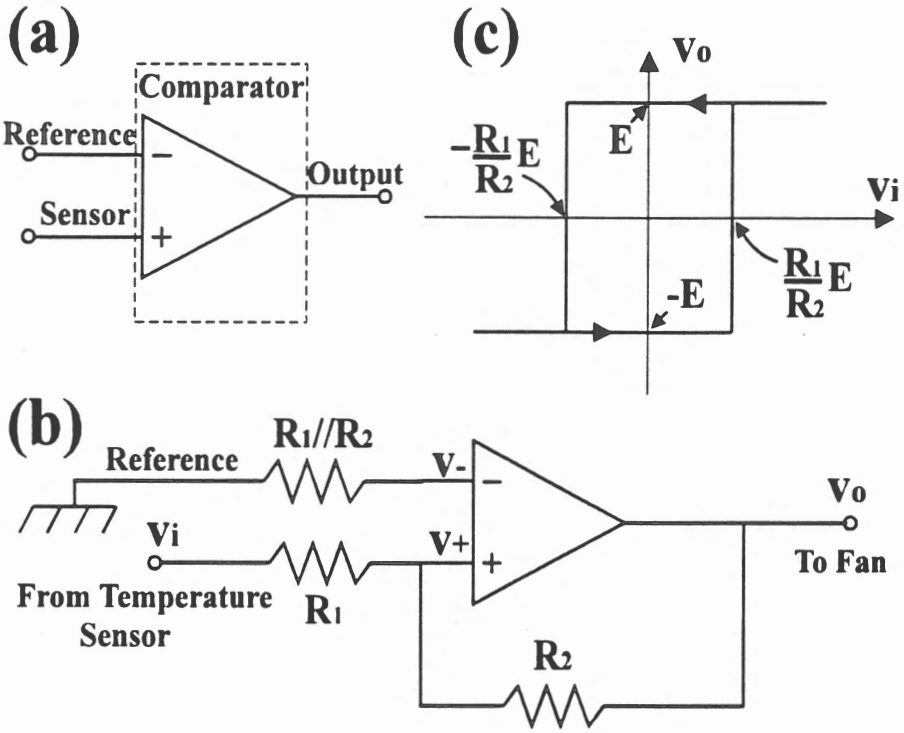


Figure 21.5: A Schmitt Trigger and Its Hysteresis Loop

21.4 Schmitt Triggers

At this point, we must stress again that op-amps are found in a huge variety of circuit types, such as rectifiers, limiters, oscillators, active filters and have several other uses, none of which there is time for here. Since we have space to select one more op-amp application, let us look at a very simple and interesting *comparator*, known as a *Schmitt trigger*. We saw earlier in the course how op-amps connected with inputs to both terminals are often used as comparators, whose output indicates the result of *testing* which of the two inputs is the higher. Imagine now that we have a temperature sensor providing one input and a voltage reference providing the other. If the actual temperature of the sensor is higher than the temperature corresponding to the reference voltage, then, in Figure 21.5a, the comparator output will be a *high* voltage. Otherwise, with the temperature below the reference, the comparator output is *low*. Now think of the op-amp output driving a cooler or fan (via a power transistor, perhaps). If the temperature were

higher than the reference, then the fan would go on, but it would stay off if the temperature were below the reference. This is a very simple sort of *thermostat*, and is just one example of a self-regulating electronic system.

If the temperature is very close to the reference, however, then the circuit of Figure 21.5a would end up switching the fan on and off rather often (indeed, this is the situation towards which the system will try to go). This may be undesirable, disturbing people nearby or eventually causing damage to the fan or switching circuitry. What is ideally needed is for the reference temperature (at which switching occurs) to fall slightly when the fan comes on, and to rise somewhat when the fan goes off, just so that the temperature does not cross the reference threshold more often than necessary.

The circuit of Figure 21.5b achieves this in a very elegant way. It is shown with a reference voltage of zero (the inverting terminal is grounded), but could be adjusted for any required reference value. Now, if v_i is well above the reference voltage, then the output will be equal to the op-amp saturation voltage, which we will call $+E$. Note that some of this output is fed back to the non-inverting op-amp terminal, so we have here an unusual situation, known as *positive feedback*. At the v_+ terminal, nodal analysis gives us

$$\frac{v_+ - v_i}{R_1} + \frac{v_+ - v_o}{R_2} = 0 \tag{21.15}$$

While the op-amp is saturated, it will not obey the Rule which states that it drives its terminals to equality under normal op-amp action. However, as v_i is reduced (because the fan is now *on*), there will come a point when the op-amp is *just* no longer saturated and we will have

$$v_+ = v_- = 0 \quad \text{with} \quad v_o = +E \tag{21.16}$$

At this instant, you can see from the nodal equation that

$$v_i = -\frac{R_1}{R_2}E \tag{21.17}$$

This is therefore the voltage to which the input voltage must fall before the op-amp comes out of saturation and the fan begins to switch off, and you will notice that it is *below* the reference voltage of zero. After this, the smallest cooling of the sensor by the fan will drive the input voltage lower, and the comparator will swiftly saturate in the opposite sense, with output $v_o = -E$. With the fan now *off*, the temperature may begin to rise, and so

does v_i, but you can see now from the nodal equation that normal op-amp action is not resumed until the moment when

$$v_+ = v_- = 0 \quad \text{with} \quad v_o = -E \tag{21.18}$$

in which case

$$v_i = +\frac{R_1}{R_2}E \tag{21.19}$$

In other words, the input voltage now has to climb *above* the reference voltage before the fan will come on again. This phenomenon, whereby the circuit is always found in one state or another and the threshold shifts to keep it in that state for longer, is known as *hysteresis*, and it implies that the system "remembers" which side of the reference voltage the input was on at the time of the last threshold crossing. Hysteresis is of great interest in a wide selection of electronic applications, and it is always characterised by *looped* graphs, such as the one in Figure 21.5c, which illustrates the behaviour of the Schmitt trigger.

With that, we come to the end of this introduction to electronics. Schmitt triggers have a number of uses of their own, including an important rôle in square-wave generation, but this is something that you are now in a position to research by yourself. At this point you may like to look back at the structure of the ideas embedded in this short course, and to reflect on the fact that modern electronics (as typified by the op-amp circuits of this chapter) is only possible because of the fascinating properties of silicon and the other semiconductor materials that we explored in the opening chapters. Connecting the two has taken us a long way through descriptions of diodes and transistors, but you are hopefully building a solid understanding of basic electronics with which to embark on more advanced studies with confidence.

AND GOD SAID:—

$$\oint \underline{E} \cdot d\underline{A} = \frac{q}{\varepsilon_0}$$

$$\oint \underline{B} \cdot d\underline{A} = 0$$

$$\oint \underline{E} \cdot d\underline{s} = -\frac{d\underline{\Phi}_0}{dt}$$

$$\oint \underline{B} \cdot d\underline{s} = \mu_0 \varepsilon_0 \frac{d\underline{\Phi}_E}{dt} + \mu_0 i$$

... AND THERE WAS LIGHT!